NBER WORKING PAPER SERIES

LEARNING VERSUS UNLEARNING:
AN EXPERIMENT ON RETRACTIONS

Duarte Gonçalves
Jonathan Libgober
Jack Willis

Working Paper 29512
http://www.nber.org/papers/w29512

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2021

For helpful comments, we thank Andreas Aristidou, Vittorio Bassi, Dan Benjamin, Aislinn Bohren, Isabelle Brocas, Gabriele Camera, Juan Carrillo, Alessandra Casella, David Cooper, Giorgio Coricelli, Mark Dean, Laura Doval, Martin Dufwenberg, Ben Enke, Dan Friedman, Mira Frick, Guillaume Frechette, Jonas Hjort, John Horton, Judd Kessler, Brian Libgober, Muriel Niederle, Kirby Nielsen, Charles Noussair, Pietro Ortoleva, Jacopo Perego, Matthew Rabin, Joao Ramos, Martin Rotemberg, Joel Sobel, Charlie Sprenger, Dmitry Taubsinky, Michael Thaler, Severine 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Learning versus Unlearning: An Experiment on Retractions
Duarte Gonçalves, Jonathan Libgober, and Jack Willis
NBER Working Paper No. 29512
November 2021
JEL No. D8,D83,D9,D91

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. Across different kinds of retractions - for instance, those consistent or contradictory with the prior, or those occurring when prior beliefs are either extreme or moderate -subjects do not fully unlearn from retractions and update approximately one-third less from them than from equivalent new information. While we document a number of well-known biases in belief updating in our data, our results are inconsistent with any explanation that does not treat retractions as inherently different. Instead, our analysis suggests that retractions are harder to process, for instance, due to the intimate reliance on conditional reasoning.

Duarte Gonçalves
Department of Economics
University College London
Drayton House
30 Gordon Street
London WC1H 0AX
United Kingdom
duarte.goncalves@ucl.ac.uk

Jonathan Libgober
Department of Economics
University of Southern California
3620 S. Vermont Ave
Los Angeles, CA 90089
libgober@usc.edu

Jack Willis
Department of Economics
Columbia University
420 West 118th Street
New York, NY 10027
and NBER
jw3634@columbia.edu

1. INTRODUCTION

Retracted information often continues to influence beliefs, even once widely discredited. Baseless rumors, false claims of politicians, mistaken earnings announcements, all tend to linger long after being revealed to be devoid of information. Perhaps the best known example is that of the study claiming a causal link between vaccination and 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, for example contributing to the uptick in measles outbreaks in the United States during the 2010s (DeStefano and Shimabukuro, 2019). For each such example there is a myriad of potential domain-specific explanations: political affiliations and motivated reasoning, inference about the motives of the original messenger or the retractor, distaste for acknowledging mistakes.[1] Yet the prevalence and diversity of failures of retractions suggests that they may reflect a common underlying cause.

Why is it so frequently easier to learn the (incorrect) information than to subsequently unlearn it? Misinformation is inevitable and influences beliefs; even in science, information thought to be true is sometimes subsequently found to be false. How do people unlearn information when it is shown to be incorrect? More generally, people can learn either from new information, or from the correction of previous misinformation. While there are many studies of the former, we know little about the latter, and whether the two work differently. Understanding how people unlearn matters for the debate about misinformation, both its harm as well as how to combat it: should we emphasize the error of the original information, or should we instead emphasize the correct alternative information? This question often also applies to information campaigns not explicitly targeting misinformation; incorrect beliefs come from somewhere.

This paper identifies and analyzes one hypothesis for this asymmetry between learning and unlearning: 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 illustrative example, imagine hearing gossip that an acquaintance committed scientific fraud, leading you to doubt their findings. Our hypothesis would imply that, upon learning that the gossip was baseless, lingering doubts about their work may remain; you may not fully "unlearn"

[1] Indeed, as we discuss at length in Section 1.1, past work in economics, political science and psychology addressing similar questions has focused on settings where these features are salient.

what you heard, in response to its retraction.

To test this hypothesis, we present an experimental design which allows us to measure (un)learning from retractions, and to compare it to learning from informationally equivalent new signals about a state. We show that, across a broad set of cases, 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 results are consistent with "information about past information" being more difficult to interpret and internalize than information that is directly informative of the state, even if the informational content is otherwise identical.

Our design is deliberately abstract, for reasons described below, and is a variation on a classic bookbag-and-poker-chips experiment. We present subjects 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, in particular 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: (1) beliefs following retractions versus beliefs without observing the retracted signal in the first place—testing whether retractions work; and (2) beliefs following retractions versus beliefs following new draws which yield identical Bayes updates (i.e. a draw of the opposite color to that which is retracted)—testing retractions versus equivalent new signals.

Our first result is that subjects fail to fully unlearn from retractions. Comparing beliefs after an earlier signal has been retracted, to beliefs after a comparable history in which 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. Moreover, retractions have less of an effect on beliefs than equivalent new information: comparing beliefs after a retraction, to beliefs after a new draw of the opposite color—two events which are informationally equivalent—beliefs update more in response to the new draw. Both results are robust across multiple variants of the experiment and hold regardless of

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

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 one-third less on average when information is a retraction (see Section 4.3.1).

This bias in updating from retractions interacts with other biases. When updating from new draws subjects (slightly) over-infer from signals and do more so when signals confirm the prior, whereas when updating from retractions they under-infer and exhibit anti-confirmation bias. Belief updating from retractions exhibits the opposite biases when compared to updating from new draws, a conclusion which is robust across all specifications. This suggests that, while new signals and retractions are informationally equivalent, these two types of information are treated in a fundamentally different manner, with retractions being much less effective at affecting beliefs.

Why are retractions less effective? 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 examine whether the diminished effectiveness of retractions could be due to three well-known behavioral biases. First, one may think that the residual effect is due to subjects having previously acted upon the retracted information and, hence, potentially internalized it. Thus, retractions being less effective could be seen as an expression of an informational endowment effect. To test this, in one treatment arm of the experiment, we randomly select subjects to only elicit their beliefs after the retractions, not after each draw. We then compare beliefs after retractions in this arm to those in the main arm, where beliefs were also elicited before the retraction. We fail to reject that retractions have the same effect on belief updating. Second, we look at whether retractions are only less effective when they refer to signals observed earlier rather than more recently. By comparing whether or not the retraction refers to the previously observed signal, we do find that retractions are slightly more effective for information which has been more recently received, but only marginally so. Third, one could conjecture that, as inference from signals has been seen to become weaker with sample size (Benjamin, 2019), perhaps retractions occurring after more information has been received are less effective. 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 them harder to interpret. Indeed our theoretical analysis illuminates this necessity by considering a general class of non-Bayesian updating models and showing that the failure of retractions cannot be seen as simply being due to other biases which do not make use of this kind of conditional reasoning which retractions necessitate. Moreover, the data further supports this conjecture: we show that subjects take 10% longer on average updating from retractions than from informationally equivalent new signals.

Finally, we leverage our design to study 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.

Our experiment is one of a small number which study how beliefs respond to *information about information*. Our findings are consistent with additional thinking costs emerging when information is presented as a retraction. That said, our analysis goes beyond simply positing that these updating frictions exist, as we exploit that our design also allows us to compare different kinds of retractions. While we see documenting that retractions are treated as distinct as a contribution in itself, we also seek to reconcile this finding with other potential explanations one may arrive at from extrapolating from other experiments and models. And, indeed, in the course of the analysis, we uncover other determinants which appear (or do not appear) to influence the effectiveness of retractions.

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 the details of any particular domain (which typically motivates interest in similar designs). As we discuss below, the closest precedent for our experiment comes from the voting literature. 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, we find this effect even without motivation. 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 well-formed. Third, we can compare retractions to other pieces of equivalent information; again, when information lends itself to a more subjective interpretation, as in the examples above, it is difficult to determine what pieces of information should be equivalent to a retraction. Fourth, our design allows us to replicate and compare with many other findings from the literature on biases in belief updating, showing that the failure of retractions to be a distinct phenomenon (e.g., not a simple case of base-rate neglect). Fifth, we are able to incentivize responses, which typically improves accuracy and reliability. While we acknowledge some of these could be addressed in other creative designs, doing all simultaneously seems important, and doing this in a simple manner seems infeasible without abstraction.

We believe the results obtained validate our focus and show the power of our design to speak to practically relevant issues. And we believe that the observation that retractions are *fundamentally less effective* has significant practical value as well. Taken together, our results provide important guidelines regarding how individuals can be expected to update about 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 paper contributes to an extensive literature in experimental economics on errors in belief updating; an authoritative and comprehensive survey on this vast literature can be found in Benjamin (2019). We replicate many of the key findings from this literature which the survey documents, such as base-rate neglect and confirmation bias.[3] Those most related to the present study are experiments on the failure of contingent reasoning. Charness and Levin (2005) was among the first papers to study the influence of such contingent reasoning in winner's curse settings, one of the first documented failures of contingent reasoning. They find that transforming the problem to one where contingent reasoning is not necessary improves participant performance. Esponda and Vespa (2014) find, in a strategic voting experiment, that subjects have difficulty extracting information from hypothetical scenarios. Martínez-Marquina et al. (2019), like Charness and Levin (2005), study the "acquire a company" game and decompose difficulties with contingent reasoning into a complexity component and a uncertainty component, finding evidence for the existence of the latter by comparing to a deterministic treatment. Enke (2020) documents that many subjects consistently fail to account for the informational content from the *absence* of a signal, suggesting a failure of contingent reasoning. We note that, with the exception of Enke (2020), these papers all examine cases where information has some instrumental use, whereas in our design information is only helpful in terms of forming beliefs in a *pure prediction* setting.[4]

Our results are consistent with added difficulties when updating from "information about information" compared to "direct information." Our theoretical framework articulates why updating from retractions necessitates the ability to perform a particular kind of conditional reasoning. While not a priori clear that this class of errors would be sufficient to detect a notable difference in our design, there is significant precedent for mistakes in contingent reasoning in probability assessments. This phenomenon is similar to what Miller and Sanjurjo (2019) refer to as the *Principle of Restricted Choice*, whereby subjects fail to condition on the data generating process behind the *source* of a piece of information. They draw a connection between this bias and the mistakes subjects make when facing the *Monty Hall Problem* (Friedman, 1998), a classic problem where subjects consistently mistakenly assess probabilities.[5] A variety of similar biases of this form have been found anecdotally.[6] We are not aware of other designs which define a natural *equivalent and*

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

[4] Of course, information is arguably instrumental in our case as well, insofar as payments depend on reports through the scoring rule. Nevertheless, our focus on a pure prediction setting is a distinguishing feature.

[5] In the Monty Hall Problem, a subject is asked to choose one of three doors, with one of the three hiding a prize. After making an initial choice, subjects learn which of the doors they did *not* select has *no* prize behind it. Subjects are then offered to switch their choice. Friedman (1998) shows that subjects—with striking consistency—choose to keep their choices, even though this has a lower probability of a prize. We note that many factors are at play in this setting, such as an illusion of control. Indeed, as far as we have been able to tell, the relevance of this problem is entirely through the Principle of Restricted Choice. See also Borhani and Green (2018), who study Monty Hall settings and characterize where an inferential bias may lead to information processing errors.

[6] See also the Bertrand Box paradox or the Boy-or-Girl Paradox for similar cases where individuals with significant

conditioning-free comparison in a single-agent, bookbag-and-poker-chips experiment, as we do in this paper.[7] This allows us to quantify precisely *how much less effective* the retraction is relative to the "conditioning-free" benchmark (i.e., direct information). In addition, our design allows us to separately identify the bias in updating from retractions from other biases in belief updating (discussed below). It also illustrate how the effectiveness of retractions varies with the kind of retraction (i.e., contradictory vs. confirmatory, or extreme beliefs vs. moderate beliefs). Difficulties in updating from retractions, by themselves, do not immediately speak to any such patterns.

Our experiment uses a bookbag-and-poker-chips experiment to argue both that (a) beliefs do not revert after a retraction, and (b) retractions are less informative than otherwise equivalent information. Past work addresses the former possibility, but typically using designs where beliefs are non-quantifiable, or that exhibit other domain-specific features (e.g., motivated beliefs). In psychology, the idea that information may have a residual impact even after it is retracted is known as the *continued influence effect*. To the best of our knowledge, this phenomenon seems relatively unexplored in economics. Johnson and Seifert (1994) illustrated that subjects would still rely upon discredited information related to the cause of a fire, and attributed this to the causal nature of the information provided. Lewandowsky et al. (2012) surveys the literature since then and highlights several possible reasons for the continued influence effect from the psychology literature. On the one hand, as mentioned above, our bookbag-and-poker-chips setup enables explicit comparisons between retractions and equivalent new information, as well as beliefs before and after the retraction. While these kinds of experiments are popular in economics, we are not aware of any past work on the continued influence effect that has utilized them (and correspondingly, our explicit comparisons of interest have been absent). Furthermore, of the four explanations given for the continued influence effect by Lewandowsky et al. (2012), none of them (in themselves) appear capable of explaining our findings.[8]

probability training can arrive at the incorrect answers.

[7] 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).

[8] 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.

Past experimental work in political science and economics seeks to quantify biases in reactions to news. However, one crucial difference with this literature is that our main treatment does not contain motivated reasoning, as may be entailed by political stories. Our results thus suggest a different mechanism than the focus of these existing papers. Substantial evidence exists that motivated reasoning is important in political settings: For instance, Thaler (2020) shows that political beliefs strongly predict that political identity influences how subjects react to information about pieces of news.[9] His design, however, explicitly seeks to eliminate inferences about source veracity, in contrast to ours.[10]

It is also worth mentioning that other biases have been documented in belief updating that we view as distinct from the phenomenon identified and which our design allows us to avoid studying. One such bias is *base-rate neglect*, whereby agents underweight the prior when updating their beliefs; see, for instance, Esponda et al. (2020). We do not vary the prior directly in order to keep our design symmetric; on the other hand, we *are* able to determine how much an initial guess influences subsequent beliefs. Another is *confirmation bias*, whereby subjects interpret information more favorably when it coincides with their initial beliefs. Such a bias is modelled theoretically by Rabin and Schrag (1999). Charness et al. (2020) experimentally study how players choose among potentially biased sources of information; however, our design features exogenous information, in contrast to theirs. We note that while we *are* able to replicate several of these results in our sample, they are not our primary focus.

2. FRAMEWORK

This Section presents our formal definition of a retraction, and includes our main framework and hypotheses.

[9] Other experimental papers in political science analyze the extent to which voters are able to rationally update beliefs include Huber et al. (2012) and Taber and Lodge (2006). These replicate certain biases in information processing, but are not about retractions per se. Angelucci and Prat (2020) study memory of news in a long-term survey and find evidence that political leanings influence which stories are likely to be remembered.

[10] For other papers studying motivated reasoning in different domains outside of politics, see Eil and Rao (2011), Mobius et al. (2013), Coutts (2019b), Grossman and Owens (2012), and Oprea and Yuksel (2020). Conlon et al. (2021) use a similar design to examine whether husbands and wives update their beliefs in the same way following information from each other versus information from strangers. A seminal theoretical contribution in this area is Brunnermeier and Parker (2005), which studies a decisionmaker who faces consumption utility, as well as *anticipatory utility* from future consumption, and shows it may be optimal for a decisionmaker to induce biased beliefs in order to maximize the latter. From a decision-theoretic perspective, Kovach (2020) studies preferences over acts and derives axioms which characterize wishful thinking, under an assumption which focuses the setting on cases where subjective expected utility holds. See Benabou and Tirole (2016) for a survey of this literature more generally.

2.1. Defining Retractions

Consider a Bayesian decision maker learning about a state of the world, which for simplicity we take to be binary, say $\theta \in \{-1, 1\}$, with prior probability p_θ; in the experiment, we take $p_\theta(1) = 1/2$. Our interest is in cases where the decisionmaker repeatedly observes independently and identically distributed signals. In particular, a signal s_t can either be *truth* or *noise*, that is,

$$s_t = n_t \cdot \epsilon_t + (1 - n_t) \cdot \theta.$$

Here, $\epsilon_t \in \{-1, 1\}$ and equals 1 with probability p_ϵ, and $n_t = 1$ whenever the signal is given by the independent "noise" ϵ_t, an event which occurs with probability p_n. When $n_t = 0$ (which occurs with with complementary probability), we have that $s_t = \theta$ and we refer to the signal as "truth"; in our experiment, we take $p_\epsilon = 1/2$ to maintain simplicity. Putting this together, the probability of observing a signal s_t given the state θ is given by

$$p_s(s_t \mid \theta) = p_n \cdot p_\epsilon(s_t) + (1 - p_n) \cdot \mathbf{1}[s_t = \theta]$$

Letting s^t denote the set of signals $\{s_1, ..., s_t\}$, we have that the posterior of a Bayesian decision maker about the state of the world θ is given by

$$p_\theta(\theta \mid s^t) = \frac{p_\theta(\theta) p_s(\theta \mid \theta)^{\#\{s_t=\theta\}} p_s(-\theta \mid \theta)^{\#\{s_t=-\theta\}}}{p_\theta(\theta) p_s(\theta \mid \theta)^{\#\{s_t=\theta\}} p_s(-\theta \mid \theta)^{\#\{s_t=-\theta\}} + p_\theta(-\theta) p_s(\theta \mid -\theta)^{\#\{s_t=\theta\}} p_s(-\theta \mid -\theta)^{\#\{s_t=-\theta\}}}$$

Definition 1. *A retraction is any signal informing the agent that a past observation s_t was noise.*

We believe this definition is in line with colloquial usage. While we specialize the definition to our main information arrival process of interest, it is straightforward to extend to the case where the data generating process does not reveal the *truth* of a state: it simply involves informing that the decisionmaker that a past signal was not reflective of the state.

In order to update beliefs as a Bayesian following a retraction, the decisionmaker must know how the retraction is generated, that is, how the retracted signal was chosen. Let τ be the period corresponding to the retracted signal s_τ. In this paper, we will focus on *verifying retractions*: these are retractions which involve first selecting a signal s_τ, and subsequently revealing to the decisionmaker whether this signal is noise or not.[11] Formally, verifying retractions are those such

[11] Note that this kind of retraction may indeed lead to the subjects learning that their past information was actually

that $\{\tau = t\}$ and $\{n_t = 1\}$ are independent events. Note that a Bayesian decisionmaker should be able to follow Bayes rule and update beliefs following retractions without any ambiguity.[12] Also note that, for a verifying retraction, we have that updating from a retraction is equivalent (as per Bayes rule) to "disregarding" the retracted signal:

$$p_\theta(\theta \mid s^t, n_\tau = 1) = p_\theta(\theta \mid s^t \setminus \{s_\tau\})$$

2.2. Belief Updating Biases

A number of studies referenced above find that subjects underweight information. If b_t is a subject's belief in period t about θ in period t, then the Bayesian belief update would yield a constant change in the $\log$ odds ratio. Let us write:

$$\log\left(\frac{b_{t+1}(s)}{b_{t+1}(-s)}\right) = \log\left(\frac{b_t(s)}{b_t(-s)}\right) + K_1\mathbf{1}[s = s_t] - K_2\mathbf{1}[s = -s_t]. \tag{1}$$

Given symmetric noise ($p_\epsilon = 1/2$), for a Bayesian under our information arrival process, $K_1 = K_2 = \log\left(\frac{2-p_n}{p_n}\right) =: K$. The literature mentioned above, however, has found that $K > K_1, K_2$, and $K_1 > K_2$ if the subject obtains additional utility when $\theta = s_t$ (and visa versa).

Various microfoundations have been proposed which would yield these biases. The first can be rationalized, for instance, by paying a "thinking cost" to choosing K, which would prevent K_ℓ from equaling K. A model of belief-based utility could induce the asymmetry in K_ℓ. To understand the theoretical difference of a retraction, note that a subject being informed that a signal were noise would not have any differential update: the first effect of a retraction is simply to "disregard" such noisy signals. However, there is also a second effect, caused by the need to infer from the retraction itself. Let $\alpha(\tau \mid s^t) = \frac{\mathbb{P}[\text{Retraction of } s_\tau | s^t, \theta=1]}{\mathbb{P}[\text{ Retraction of } s_\tau | s^t, \theta=-1]}$. With symmetric noise, if signal s_τ is retracted, the Bayesian update should be:

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

where $\eta_t := \sum_{\ell=1}^t s_\ell$. Then, for a retraction, the log odds update is now:

accurate; however, these signal realizations are (in principle) degenerate at certainty, and so we do not use any such reports in our analysis in any significant way. Our companion paper studies a version of this design which *does* allow us to study such positive verification directly.

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

$$\log\left(\frac{b_{t+1}(s)}{b_{t+1}(-s)}\right) = \log\left(\frac{b_t(s)}{b_t(-s)}\right) + K_1\mathbf{1}[-s \text{ retracted}] - K_2\mathbf{1}[s \text{ retracted}] + \log(\alpha(\tau \mid s^t)), \quad (2)$$

which is identical to our previous expression, except for the $\log(\alpha(\tau \mid s^t))$ term. Now, $\alpha(\tau \mid s^t) = 1$ (and hence $\log(\alpha(\tau \mid s^t)) = 0$) for all verifying retractions. The purpose of our experiment is to understand whether subjects treat $\alpha(\tau \mid s^t) = 1$. We also wish to understand whether the *reasons* for any departure are due to the same biases which would lead to any of those biases identified in prior work.

One final comment prior to the analysis: In our analysis, we will be interested in both belief levels, as well as $\log$ odds, and think each measurement has merits. Nevertheless, studying whether $\log(\alpha(\tau \mid s^t)) = 0$ in (2) to determine whether retractions are treated differently does not require the decisionmaker to be a Bayesian. This observation is due to a result of Cripps (2019). The author shows that as long as the decisionmaker's rule is "divisible" (which roughly states signals are treated as exchangeable), the updated belief must be found via a transformation of the prior, Bayes rule applied to the transformation, and an inverse transformation; updating rules satisfying this requirement are commonly used in experimental work (e.g. Angrisani et al., 2019). Thus, under any updating rule in this general class, one has that the reported belief b_t is a function of the Bayesian belief, and that b_{t+1} is the inverse of the Bayesian update of the transformed belief; that is, we have for some monotonic f, belief updates can be determined via the following relationship:

$$\log\left(\frac{f(b_{t+1}(s))}{1 - f(b_{t+1}(s))}\right) = \log\left(\frac{f(b_t(s))}{1 - f(b_t(s))}\right) + K_1\mathbf{1}[-s \text{ retracted}] - K_2\mathbf{1}[s \text{ retracted}] + \log(\alpha(\tau \mid s^t)), \quad (3)$$

The key observation from this expression is that, since $\log(\alpha(\tau \mid s^t)) = 0$, no model of "as-if" Bayesian updating would explain any difference due to retractions. Indeed, this observation would hold for *any* f—including those that may rationalize under-updating, for instance. We record this result as follows:

Proposition 1. *Consider any (possibly non-Bayesian) updating rule, which emerges as the result of "Bayesian updating under a transformation." Any such decisionmaker would have an equivalent* $\log$ *odds update following a retraction of a signal* s *compared to a new signal* $-s$*, provided* $\alpha(\tau \mid s^t) = 1$.

Even when focusing on levels, however, this result strengthens the interest in our design as stepping beyond the normal boundaries of Bayesian updating. Any differential effect we find would not be consistent with any updating rule that satisfies exchangeability. This will be helpful in distinguishing our results from similar analyses in the literature.

2.3. Hypotheses

The purpose of this paper, simply put, is to understand patterns in $\alpha(\tau \mid s^t)$. 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 hypothesis:

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

It is worth emphasizing 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 some other departure from Bayesian updating yielding a difference with new information, leading to (b) without (a). As we do not seek to test Bayesian updating per se, we also do not wish to imply that these two effects are equivalent.

Despite the fact that our anecdotal evidence is highly suggestive that this hypothesis should hold, it should not be immediately apparent that we should be guaranteed to find this in our design. A priori, it could very well be that our setting is *too* stripped down and has therefore eliminated whatever aspects of these applications is responsible for the ineffectiveness of retractions. If subjects are always more likely to make mistakes following more signals, for instance, then a retraction may be easier to internalize since it suggests that a subject only need to consider a smaller number of signals.

The substance of the hypothesis, then, is that we attribute a distinctive effect to retractions themselves relative to new signals. One expects, from prior work, that $K_\ell < K$ when regressing log odd updates on log odd initial beliefs. However, our focus is not testing whether or not subjects abide by Bayesian updating rigorously; we expect to find various biases in our sample. In contrast, while acknowledging that subjects may exhibit biases relative to the Bayesian standard, our conjecture is that belief updating from retractions is *distinctively* less accentuated.

Note that this hypothesis does assume that subjects understand a basic property of how retractions are generated; specifically, as explained above, a retraction suggests a signal should be "deleted," and not that it is actually evidence for the opposite. While certainly debatable (and in any case not relevant for our findings), we do believe this is more in line of what a retraction actually is. In many instances, if a study uses fabricated data, then the inference upon learning this should only be that there was no informational content to the results—even if a correctly done study would have worked. It is worth noting that a priori there is no way of knowing whether subjects *interpret* retractions in this alternative way, even if we present a retraction to them as a signal deletion. Naturally, we strove to provide subjects with clear and concise explanations about the data-generating process that would lead them to correctly understand the meaning of retractions. However, if subjects misinterpreted what we meant by a retraction, then it would certainly be possible for this effect to override the intuition and invalidate the hypothesis. In this sense, if anything, the possibility of misinterpreting the data-generating process in this way makes it more likely that the hypothesis is falsified by the data.

This discussion also highlights how subtle and complex belief updating can be. In turn, the complexity of belief updating is also often used in order to explain the emergence of other commonly studied biases commonly. Insofar as retractions might be harder to internalize, this suggests our second hypothesis:

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

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

We then test for the above-mentioned explanation as to why retractions may work differently from new signals. 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 question is whether it depends on the sequential nature of information arrival implied by our framework. Does the fact that subjects are required to report their beliefs after every signal make their beliefs harder to move, akin to an informational endowment effect?

Hypothesis 3 (Retracting internalized signals)**.** *The effect of retractions on belief updating is weaker when agents acted upon the observed signals.*

Relatedly, insofar as our experiment involves dynamic information arrival and correspondingly may involve dynamics in terms of the difficulty of updating, one might expect there to be some interaction with the timing of retractions (as well as the timing of the signal that is retracted). In line with existing evidence for the well-known serial-position effects in belief updating such as the recency effect—the tendency to overweight recent signals relative to others—a possible conjecture is that retractions would lead to (more effective) unlearning of past information when such 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 mechanism that could be at play in rendering retractions ineffective is that retractions could potentially 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, (log-odds) 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 has a lower impact on belief updating, which we take as a conjecture:

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. Our conjecture is that the complexity inherent to 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, insofar as 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 first set of hypotheses posit that subjects may display something similar to an "endowment effect for past signals." That is, a decisionmaker observing a particular signal s_t, having incurred a cost to update their beliefs, would be resistant to deleting or internalizing it. In this light, several of our subsequent hypotheses hinge on what the shape of these costs may be (e.g., is it even more costly to update after some costs have been incurred). Accordingly, a natural question is whether the effectiveness of retractions depends only on the beliefs themselves, or how the decisionmaker has used information, or preferences over beliefs. While these elements are (to varying degrees) known to influence Bayesian updating, as Proposition 1 makes clear, this is orthogonal to the question of whether it influences the effectiveness of retractions.

2.4. Outline for Analysis of the Hypotheses

We describe our basic design first, which describes the method of drawing retractions which subjects faced during the experiment. Subsequently, we describe additional details of each round of the experiment, including exact variations considered and other details. Subjects were provided full information regarding how observations would be drawn and compensation would be provided. See Appendix B for the instructions as presented to the subjects in the experiment.

3. EXPERIMENTAL DESIGN

The purpose of the experiment is to study how people learn from retractions. There are two main questions. First, are retractions effective? That is, once information is retracted, do people behave as if they never received the information to begin with? Second, is learning from retractions inherently different from learning from new signals? While answering these two questions, we also provide evidence on when and why updating from retractions is different.

We purposefully test these questions in an abstract setting, one of drawing colored balls from urns. Existing experiments on retractions are in settings where many possible mechanisms are at play, for example the retraction of politically polarizing newspaper articles, where motivated reasoning and complicated inference regarding the strategic incentives of others may influence belief updating. We wish to test whether there is an underlying, cognitive difference in processing retractions, free of such confounding mechanisms.

The basic data generating process in the experiment—described in the previous Section—is implemented as described below.

Subjects play multiple *rounds* of the experiment. At the start of each round, a state is drawn, which we refer to either as "yellow" or "blue," with each state being equally likely. The state refers to the color of a particular ball, which we refer to as the *truth ball*. Subjects are told that the truth ball is placed in a box which additionally contains *noise balls*; these are yellow and blue in equal proportion.

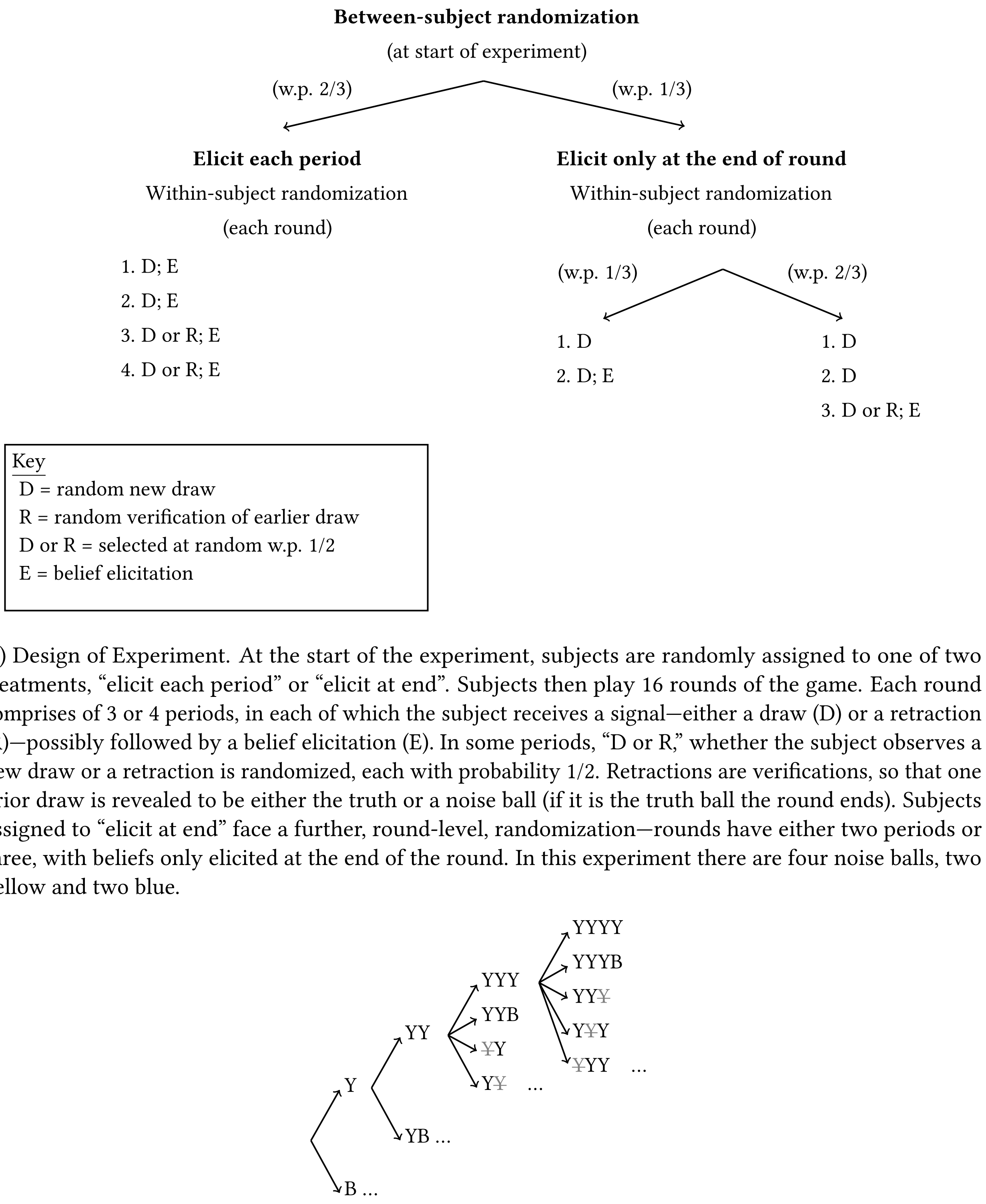
Rounds consist of multiple *periods*. In each period, subjects receive a new piece of information: either a new signal, or a retraction. A *new signal* corresponds to a ball being drawn from the box, with replacement, and subjects being told the color of the ball, but not whether it was the truth ball or a noise ball. A *retraction* corresponds to informing the subject that a past draw was a noise ball, in which case it (generally) contained no information regarding the state. Our experiment focused on verifying retractions.[13]

At the end of each period—that is, after each new piece of information—subjects report their belief regarding the probability of the truth ball being blue or yellow with these reports incentivized as detailed below. Comparing beliefs after a given ball draw is retracted to beliefs under histories in which the retracted ball draw was never made to begin with allows us to test whether retractions are effective. Comparing changes in beliefs in response to retractions to changes in beliefs in response to equivalent ball draws, allows us to test whether learning from retractions is different from learning from new signals. Indeed, a key aspect of our design is that a new draw of one color is informationally equivalent, for a Bayesian, 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.

We ran the experiment on Amazon Mechanical Turk (henceforth MTurk) in June 2020.[14] The

[13] As we mentioned, in ongoing research we also ran a version of this experiment using falsifying 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 $\log(\alpha(\tau \mid s^t)) = 0$. These results are available from the authors upon request.

[14] All our subjects were recruited in this way. By now, using MTurk itself does not appear to be a distinguishing



(a) Design of Experiment. At the start of the experiment, subjects are randomly assigned to one of two treatments, "elicit each period" or "elicit at end". Subjects then play 16 rounds of the game. Each round comprises of 3 or 4 periods, in each of which the subject receives a signal—either a draw (D) or a retraction (R)—possibly followed by a belief elicitation (E). In some periods, "D or R," whether the subject observes a new draw or a retraction is randomized, each with probability 1/2. Retractions are verifications, so that one prior draw is revealed to be either the truth or a noise ball (if it is the truth ball the round ends). Subjects assigned to "elicit at end" face a further, round-level, randomization—rounds have either two periods or three, with beliefs only elicited at the end of the round. In this experiment there are four noise balls, two yellow and two blue.

(b) A trimmed event tree of the histories which subjects could see, by period, in "elicit each period" treatment. The 'trimming' is that we only expand the top branch after each period in the figure. Y represents a yellow ball, B represents a blue ball, strike-through represents a retraction of an earlier draw (we exclude verifications which reveal the truth ball, as the round ends at that point).

Figure 1: Experimental design

experiment has several sub-treatments, described in detail below, focused on verifying retractions. Subjects were provided all information regarding how observations would be drawn and how they would be compensated. See Appendix B for the instructions as presented to the subjects.

In the rest of this section, we describe additional details of each round of the experiment, which are also explained in Figures 1 and 2.

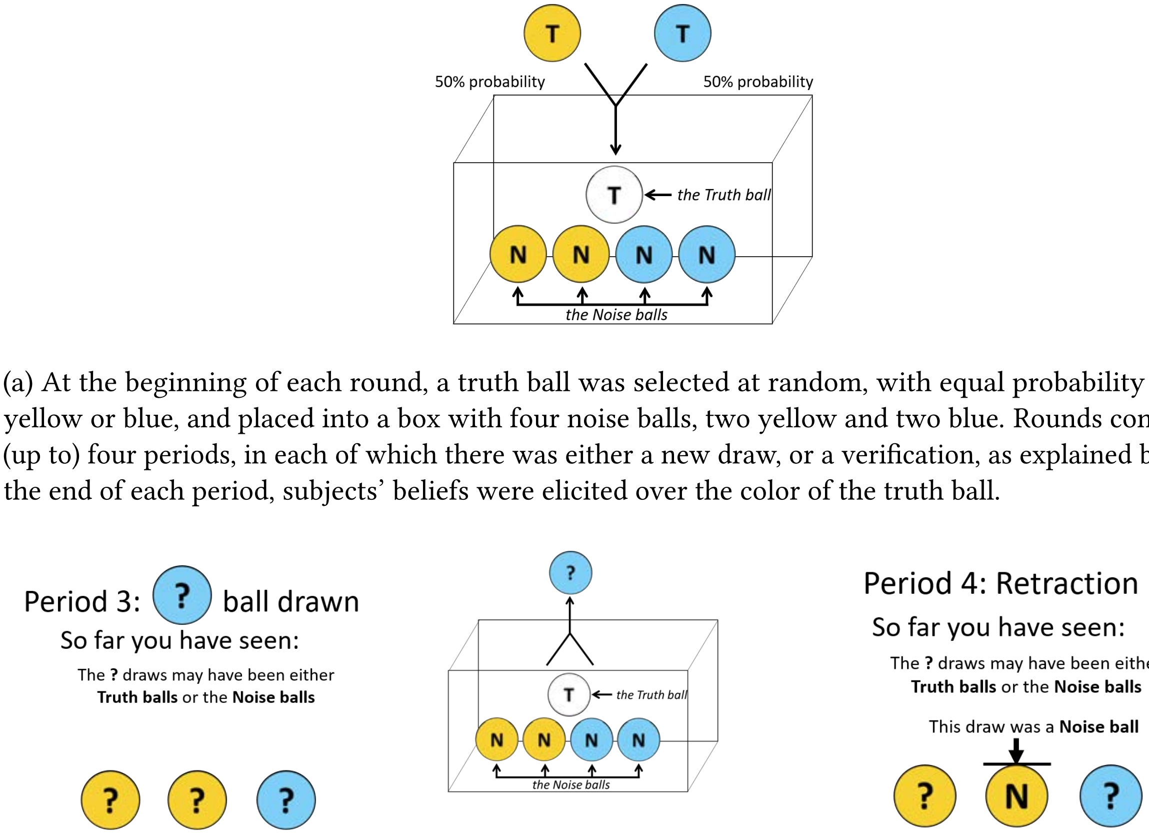
3.1. Baseline Setup

In the main treatment, each round has four periods, with beliefs elicited at the end of each round. The sequence of events is the following:

- At the start of the round a truth ball is chosen at random (either yellow or blue with equal probability) and placed into the box with two yellow noise balls and two blue noise balls (corresponding to $p_n = 4/5$ in the information arrival process).
- In each of periods one and two, the subjects observe a draw from the box, with replacement, as described above. They are told the color of the ball but not whether it is the truth ball or a noise ball.
- In each of periods three and four, subjects are provided a new draw (as above) with probability 1/2, or a verifying retraction with complementary probability. A verifying retraction worked by choosing one of the prior draws at random and informing the subject whether it was a noise ball—a "retraction"—or a truth ball. If it was revealed that the ball was a truth ball, the round was stopped, as at that point the state was fully revealed. The probability that they would receive each of these signals was determined independent across periods.

A summary of the explanatory visuals shown to subjects is given in Figure 2 and the full instructions of the experiment can be found in Appendix B.1. Beliefs were reported using a slider, which

feature; however, our scale is somewhat larger than a typical study, and our design replicates certain documented phenomenon from lab experiments using MTurk. In our case, the plethora of possible belief paths implies that a lab experiment would be subject to a nontrivial amount of additional sampling noise, in that certain paths may not be observed with sufficient frequency given a smaller participant pool. Given existing evidence, it is hard to document cases in the experimental economics literature where the results were driven by the use of venue; see, for instance, Landier et al. (2020), Martínez-Marquina et al. (2019), and Arechar et al. (2018). However, we acknowledge, as Snowberg and Yariv (2020) note, there may nevertheless be a tradeoff between the noise in the MTurk participant pool and other commonly used participants such as university students. We determined that for our design, the ability to easily recruit additional subjects outweighed the costs. While perhaps less novel in terms of economic implications, we nevertheless believe that these replications are important, given the anticipated growth of this platform in future work.



(a) 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) In periods where there was a new draw, a ball was drawn from the box (with replacement), and the color of the drawn ball was disclosed, but whether it was the truth ball or a noise ball was not. The history of the round was displayed throughout. In periods where there was a verification, 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.

Figure 2: Summary of experimental visuals

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.

Comparing beliefs after a retraction in period 3, to the equivalent beliefs in period 1, allows us to test if retractions are effective (Hypothesis 1a). Since a retraction of one color in period 3 is equivalent, from a Bayesian perspective, to a new draw of the other color in period 3, comparing updating under each of these events allows us to test whether learning from retractions is somehow different from learning from new signals (Hypothesis 1b). Finally, comparing updating in period 4 after a retraction in period 3 to updating after a new draw in period 3 (and also to updating in period 2) allows us to test whether retractions affect subsequent updating (Hypothesis 7).

3.2. Single-Elicitation Design Treatment

Our experiment also features an across-subject treatment. At the start of the experiment, each subject is randomly allocated to one of two treatments. With 2/3 probability they are allocated to the *intermediate-elicitation treatment*, exactly as described above. With 1/3 probability, they are allocated to the *single-elicitation treatment*, which differed in several ways.

In the single-elicitation treatment, subjects observe two signals with probability $1/3$ and three signals with probability $2/3$. Most importantly, beliefs are only elicited at the end of each round, not at the end of each period. The sequence of events, summarized in Figure 1, is as follows:

- At the start of the round a truth ball is chosen at random.
- The first two periods are always new draws. Beliefs are not elicited after period 1. With probability 1/3, beliefs are elicited after period two, after which the round ends. With probability 2/3, beliefs are not elicited after period two and the round continues to period 3.
- If reached, period 3 is a new draw with probability 1/2 and a (verifying) retraction with complementary probability.[15] Beliefs are elicited at the end of the period, after which the round ends.

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.

This variant of the design tests whether requiring that subjects provide reports in *every* period affects the efficacy of retractions (Hypothesis 3). One hypothesis is that discarding information which has already been internalized or acted upon is difficult. If so, we may expect retractions to be more effective when beliefs have not already been elicited. This treatment has the obvious downside that we have substantially less data to populate belief paths, and so more subjects are needed to obtain similarly precise estimates. Furthermore, since we only wanted to obtain *one* report, we needed to vary the number of periods which comprised a round.

[15] This motivates the asymmetry in whether subjects would observe two or three signals; this ensures we have an equal number of subjects reporting after two new draws, after three new draws, and after two new draws following a retraction.

3.3. Other Experimental Design Details

Subjects needed to answer comprehension questions in the instructions correctly in order to proceed with the experiments. The questions summarized the key points the subjects needed to understand. We also asked additional questions on mathematical ability, which were incentivized by providing a 50 cent reward for every question answered right. Lastly, subjects were given two rounds of "practice" to familiarize themselves with the interface. These rounds simply showed the subjects examples of what they would do during the experiment, and were not incentivized.

We incentivized subjects to report truthfully using a binarized scoring rule (see Hossain and Okui (2013) and Mobius et al. (2013)). It is well known that elicitation methods are often difficult for participants in experiments to understand or appreciate; in the interest of transparency, we displayed in the interface exactly how a report would correspond to payoffs, which subjects would be able to see if they wished to. This avoided the need for subjects to do any computation on their own, since they would see exactly what the payouts would be as a function of the truth ball's color. However, in the instructions, we simply mentioned to subjects that we believed it would be in their best interest to report the truth, and did not require them to understand details of the payment scheme directly (or reasons behind its incentive properties). In order to determine the agent's payments, we used one of their reports within a single randomly selected period.

4. RESULTS

There are four sets of principle results. First, before turning to the contribution of our paper, in Section 4.2 we show that updating from new draws is similar in our experiment to what has been found in the existing literature, to validate our experimental setting. Second, in Section 4.3, we turn to updating from retractions: do retractions work?; do people update differently from retractions versus new draws?; and how do retractions interact with existing deviations from Bayesian updating documented in ball-draw experiments? (Hypotheses 1-2). Third, in Section 4.4, we examine possible mechanisms underlying retraction ineffectiveness, addressing the question of why retractions fail (Hypotheses 3-6). Finally, in Section 4.5, we test whether retractions affect *subsequent* updating (Hypothesis 7).

4.1. Methodology Overview

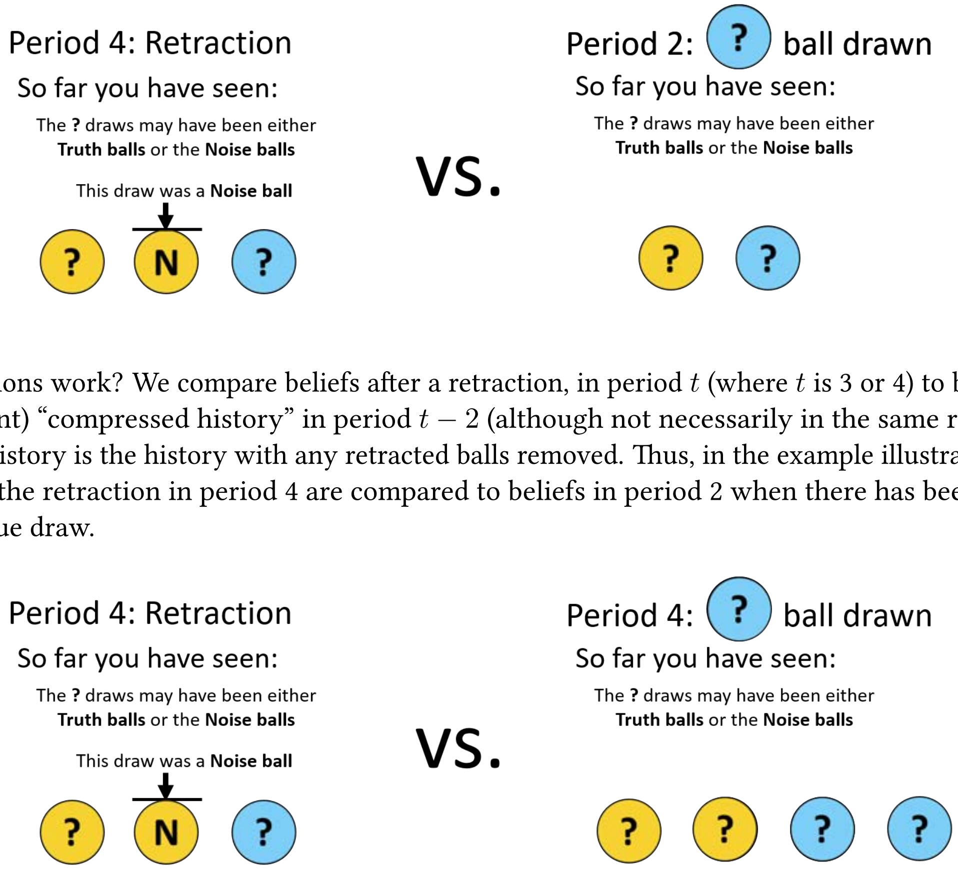
Our analysis leans on the simplicity of our experimental design to make the analysis as non-parametric as possible; however, at times we add slightly more structure, following specifications used in the literature on belief updating. We broadly perform two distinct kinds of comparisons, explained visually in Figure 3, which correspond directly to parts a and b of Hypothesis 1, but which we also use for tests of subsequent hypotheses:

- *Tests of unlearning*: Are subjects' beliefs after seeing a retraction the same as if the retracted signal had never been observed in the first place?
- *Comparisons to new information*: Do subjects treat retractions as having similar informational content as equivalent (in terms of Bayesian belief updates) new information?

When analyzing the first question, tests of unlearning, we compare the belief reports themselves, as only the level comparisons are relevant. By contrast, the second question relates to how beliefs *move* in response to retractions; for this, we report both differences in log odds, as well as levels. Levels has the advantage that extreme beliefs, near 0 or 1, 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.

While the experimental design makes our two comparisons of interest very simple, we need to aggregate them across multiple histories, both for simplicity of exposition and for statistical power. Before presenting our analysis in detail, we first outline how we do this aggregation, and how we are pinpointing the identifying variation of interest, as explained visually in Figure 3. For our tests of unlearning, when aggregating across histories we use what we refer to as a *compressed history*; this involves removing any retracted ball draws, as if those events had never occurred (so, for instance, a history of *blue-blue-retraction* would be equivalent to *blue*). Note that while the signal order does matter, we do *not* include a timestamp on each signal draw for these fixed effects. We include fixed effects for compressed history in our analysis, and hence compare beliefs within the same compressed history: as if a ball draw was never received, compared to after it was retracted. We interact this with whether the signal is yellow or blue.

For the comparisons to new information (i.e., the second bulletpoint above), we include fixed effects at the level of the *lagged history*. Here, the lag refers to the history from one draw before, which implies we control for the initial beliefs prior to the observation of the signal of interest.


Period 4: Retraction
So far you have seen:
The ? draws may have been either Truth balls or the Noise balls
This draw was a Noise ball
?
N
?
VS.
Period 2: ? ball drawn
So far you have seen:
The ? draws may have been either Truth balls or the Noise balls
?
?
Period 4: Retraction
So far you have seen:
The ? draws may have been either Truth balls or the Noise balls
This draw was a Noise ball
?
N
?
VS.
Period 4: ? ball drawn
So far you have seen:
The ? draws may have been either Truth balls or the Noise balls
?
?
?
?

(a) Do retractions work? We compare beliefs after a retraction, in period t (where t is 3 or 4) to beliefs after the (equivalent) "compressed history" in period $t-2$ (although not necessarily in the same round). The compressed history is the history with any retracted balls removed. Thus, in the example illustrated, beliefs elicited after the retraction in period 4 are compared to beliefs in period 2 when there has been a yellow and then a blue 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, but necessarily in a different round. Thus, in the example illustrated, beliefs elicited after the retraction of the yellow ball in period 4 are compared to beliefs elicited in period 4 when the history through period 3 is the same, but then period 4 is a draw of a blue ball.

Figure 3: Illustrative examples to explain the two main identification approaches

Put differently, this dummy groups histories which would lead to the same Bayesian belief updates before observing the subsequent signal (which may either be a retraction or a new draw). In addition, we include variables for whether the signal is evidence for yellow or blue ($+1$ if yellow, -1 if blue), whether the signal observed is a retraction, and an interaction term (i.e., whether the signal observed was evidence of yellow or evidence of blue). Our main interest for the comparison to new information is in the sign of the interaction term, which capture how subjects interpret information conveyed by a retraction in comparison to a new equivalent signal. Insofar as we have no reason to suspect a retraction to be interpreted more favorably as evidence for one color over the other, we expect this dummy to be a 0 (given that we are controlling for history, and the symmetry of our design). However, a negative coefficient on the interaction term suggests beliefs move less when a given signal is a retraction, controlling for history and the sign of the signal.

4.2. Preliminary Observations on Belief Paths

As a first step, in part as a test of validity of experimental setting, we examine the belief paths of subjects when they are not shown retractions. In the absence of a retraction, the design is very similar to many others surveyed by Benjamin (2019). To start with, 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.

One concern about our design is that the complexity would make it difficult for subjects to understand the instructions. Nevertheless, we do find that subjects tend give reports that are consistent with predictions one would expect with the literature. Figures 4 and 5 in the Online Appendix 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, despite the aforementioned lower effect. We note that while on a substantial number of occasions subjects appear to misinterpret information by giving reports in the opposite direction one would expect,[16] the overwhelming majority of subjects do not make this mistake (or at least do not make it consistently).

We present Grether-style (Grether, 1980) log-odds regressions—the workhorse model of analysis in this literature—in Table 1, as well as Table 11 in the Online Appendix, 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 prior retraction in the round (so only new draws):

$$l_t = \beta_0 + \beta_1 \cdot l_{t-1} + \beta_2 \cdot s_t \cdot K \tag{4}$$

$$\text{and} \qquad l_t = \beta_0 + \beta_1 \cdot l_{t-1} + \beta_2 \cdot s_t \cdot K + \beta_3 \cdot s_t \cdot K \cdot c_t \tag{5}$$

where t is the period, l_t is the log-odds of the beliefs reported at t, 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.

[16] Subjects' change in belief reports goes against the signal 19.5% of the time.

	(1) l$_t$	(2) l$_t$	(3) l$_t$
Prior (l_{t-1})	0.883***	0.838***	0.825***
	(0.0233)	(0.0274)	(0.0298)
Signal (s_t)	1.321***	1.042***	1.042***
	(0.0463)	(0.0611)	(0.0642)
Signal Confirms Prior ($s_t \cdot c_t$)		0.561***	0.583***
		(0.110)	(0.111)
Signal × Prior ($s_t \cdot l_{t-1}$)			-0.229**
			(0.0908)
Prior × Signal Confirms Prior ($s_t \cdot l_{t-1} \cdot c_t$)			0.274*
			(0.148)
Observations	11739	11739	11739
R-Squared	0.425	0.427	0.429

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1: Updating from new draws. This table represents updating from standard new ball draws. The sample consists in subjects in the intermediate elicitation 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. 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 4, 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 $t-1$.

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. In contrast, in the most parsimonious of our regressions, we find this signal to be 1.321, indicating 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 1.042 (not significantly different from 1), and

$\beta_3 > 0$. Together, this suggests that our subjects 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. Thus, while most of the studies report $\beta_2 < 1$, strict over-inference resulting from confirmatory information—that is, $\beta_2 + \beta_3 > 1$— has been previously documented (e.g. Charness and Dave, 2017).

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 these findings we mentioned are essentially what one would expect based on the literature. It also suggests that, since we do find these biases in the "new information" treatment, any additional departure due to retractions cannot be attributed to explanations that do not use the nature of the information source.

We also run the following specification, interacting prior beliefs with both the signal and the signal interacted with whether it is confirmatory.

$$l_t = \beta_0 + \beta_1 \cdot l_{t-1} + \beta_2 \cdot s_t \cdot K + \beta_3 \cdot s_t \cdot K \cdot c_t + \beta_4 \cdot l_{t-1} \cdot s_t + \beta_5 \cdot l_{t-1} \cdot s_t \cdot c_t \tag{6}$$

Under this specification, $\beta_4 < 0 < \beta_5$ represents belief entrenchment, i.e. resisting revising beliefs according to information that goes against prior; this phenomenon is more expressive the more extreme the prior is.

To summarize, in our analysis of this data, we do not see any consistent departure from the prior literature on belief updating. Subjects do display under reaction to new information in general and typically move in the correct direction. We do not find any significant departures from the main conclusions of Benjamin (2019), and therefore do not have strong reasons to suspect our results are driven by, for instance, the choice of venue.

4.3. 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.

	(1) b$_t$	(2) b$_t$	(3) Δ b$_t$	(4) Δ b$_t$
Retraction (r_t)	-0.000492	-0.00367	-0.00406	-0.00353
	(0.00341)	(0.00430)	(0.00383)	(0.00366)
Retracted signal ($r_t \cdot s_t$)	-0.0220***	-0.0322***	-0.0285***	-0.0264***
	(0.00305)	(0.00430)	(0.00383)	(0.00375)
Signal (s_t)				0.0762***
				(0.00262)
Compressed history FEs	Yes	No	No	No
Lagged history x Sign of Signal FEs	No	Yes	Yes	No
Lagged history FEs	No	No	No	Yes
Observations	17591	9074	9074	9074
R-Squared	0.154	0.255	0.126	0.122

Standard errors 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)? The sample consists in subjects in the intermediate elicitation 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 a verification 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. $r_t \cdot s_t$ is the retracted signal in round t (+1 or -1). The regression includes fixed effects for the compressed history of draws. Columns (2) and (3) 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 validations or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the history of the previous period interacted with the signal. In column (2), the outcome is the beliefs in period t, b_t. In column (3) and column (4), the outcome is the first difference in beliefs.

4.3.1. Failure to "Unlearn" and Retractions Versus New Signals

Our first result, and the key finding of the paper, is the empirical support of Hypothesis 1: retractions are ineffective, in that (1) retracted signals are not fully disregarded, and (2) beliefs are less responsive to retractions than new signals. The results on this are presented in Table 2. We run three tests, which look at how retractions both change the absolute magnitudes of the belief reports, as well as the change in the belief reports in response to the new information. The former can be seen in the first two columns, and the latter can be seen in columns three.

Our basic strategy for identifying the effects of retractions on belief updating is simple, as summarized in figure 3. In order to aggregate these simple effects across different histories, let us introduce some notation. Let H_t be the history up to time t, that is, the set of all the draws observed as well as which ones were retracted, fixing the order. As mentioned above, we call the compressed history $C(H_t)$ the set of all draws observed *excluding* any retracted ones, but keeping the order fixed. For example, if at period 4 the history is $H_4 = (s_1, s_2, n_2 = 1, s_4)$—with $n_2 = 1$ corresponding to a retraction of the second draw—then the compressed history $C(H_4)$ is given by (s_1, s_4). We will denote F_{H_t} and $F_{C(H_t)}$ the fixed effects associated with the history up to time t and those associated with the compressed history.

The first of these tests assesses whether updating from retractions is equivalent to deleting the retracted piece of information. We do this by estimating the following equation:

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

where r_t is an indicator variable for the signal being a retraction, and s_t gives the direction implied by the signal or retraction observed.[17] Controlling for the compressed history allows us to compare, for example, the beliefs after observing $\{s_1, s_2, n_2 = 1\}$ to those reported when only signal s_1 was seen.

We then look at whether belief updating is the same for a retraction of a signal s as for an informationally-equivalent new signal $-s$. We do this by estimating two regressions. First:

$$b_t = \beta_0 + \beta_1 \cdot r_t + \beta_2 \cdot r_t \cdot s_t + \beta_3 s_t + F_{H_{t-1} \times \{s_t\}}, \tag{8}$$

where $F_{H_{t-1} \times \{s_t\}}$ denote the fixed effects associated with the lagged history interacted with the direction of the signal or retraction observed. Thus, simply from changing the fixed-effects, we can now compare beliefs reported after $\{s_1, s_2, n_2 = 1\}$ to those reported after $\{s_1, s_2, s_3 = -s_2\}$.

The third specification also compares retractions to new signals, but we instead consider the *change* in beliefs by estimating the following equation:

$$\Delta b_t = \beta_0 + \beta_1 \cdot r_t + \beta_2 \cdot r_t \cdot s_t + \beta_3 s_t + F_{H_{t-1} \times \{s_t\}}. \tag{9}$$

[17] To be precise, if a signal with value s is retracted, then $s_t = -s$.

The key finding is that β_2, across all of the specifications we study, is negative—while beliefs do move following retractions, their effectiveness is dampened, since they move less strongly in the direction of the signal. This implies not only that retracted signals are not fully disregarded (column (1)), but also that providing informationally-equivalent new signals is more effective in inducing changes in beliefs (columns (2) and (3)). In short, retractions are treated *differently*, and in particular as if they were less informative.

To help quantify our results in straightforward terms, we also provide a back-of-the-envelope calculation showing 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 looking at the ratio between the coefficient on the retracted signal—the interaction term between the signal and the retraction variables—and the coefficient on the signal variable itself. This particular calculation uses lagged history fixed effects, rather than lagged history interacted with signal fixed effects, in order to determine a meaningful estimate on the coefficient of s_t; reassuringly, the estimate of the coefficient on $r_t \cdot s_t$ remains robust to these less restrictive fixed effects. We also performed 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, and we consistently find that beliefs update around $1/3$ from retractions relative to new signals.

4.3.2. Variation Across Paths: Retractions Accentuate Biases in Updating

Having illustrated that retractions are treated differently, with beliefs reacting less on average, we then seek to determine when this effect is relatively more or less pronounced. Specifically, we now adopt a similar regression specification to prior bookbag-and-poker-chips experiments, and examine heterogeneities in the relative effect of retractions across different belief paths.[18] Specifically, we run variants of the following regression, which uses log odds as the dependent variable to replicate existing papers on deviations from Bayesian updating:

$$\begin{aligned} l_t = \beta_0 + \beta_1 \cdot l_{t-1} + \beta_2 \cdot s_t \cdot K + \beta_3 \cdot s_t \cdot K \cdot c_t + \beta_4 \cdot l_{t-1} \cdot s_t + \beta_5 \cdot l_{t-1} \cdot s_t \cdot c_t + \\ + r_t \cdot [\gamma_0 + \gamma_1 \cdot l_{t-1} + \gamma_2 \cdot s_t \cdot K + \gamma_3 \cdot s_t \cdot K \cdot c_t + \gamma_4 \cdot l_{t-1} \cdot s_t + \gamma_5 \cdot l_{t-1} \cdot s_t \cdot c_t] \qquad (10) \end{aligned}$$

[18] These regressions involve weighting the error terms based on the round they are emerging from. This is because certain rounds—particularly those involving a mix of blue and yellow draws—will mechanically be oversampled by our design. Our weighting adjusts for this in order order to ensure this oversampling does not yield bias in the direction of the most likely paths.

	(1)	(2)	(3)
	l$_t$	l$_t$	l$_t$
Prior (l_{t-1})	0.883***	0.840***	0.820***
	(0.0371)	(0.0438)	(0.0497)
Signal (s_t)	1.417***	1.185***	1.163***
	(0.0628)	(0.0834)	(0.0937)
Retraction (r_t)	-0.0274	-0.0135	-0.000210
	(0.0322)	(0.0326)	(0.0423)
Retraction × Prior ($r_t \cdot l_{t-1}$)	-0.0793	-0.00832	-0.0188
	(0.0534)	(0.0637)	(0.0681)
Retraction × Signal ($r_t \cdot s_t$)	-1.027***	-0.656***	-0.693***
	(0.0798)	(0.114)	(0.122)
Signal Confirms Prior ($s_t \cdot c_t$)		0.475***	0.534***
		(0.156)	(0.165)
Retraction × Signal Confirms Prior ($r_t \cdot s_t \cdot c_t$)		-0.809***	-0.817***
		(0.210)	(0.210)
Signal × Prior ($s_t \cdot l_{t-1}$)			-0.234
			(0.159)
Prior × Signal Confirms Prior ($s_t \cdot l_{t-1} \cdot c_t$)			0.316
			(0.248)
Retraction × Signal × Prior ($r_t \cdot s_t \cdot l_{t-1}$)			0.209
			(0.205)
Retraction × Prior × Signal Confirms Prior ($r_t \cdot s_t \cdot l_{t-1} \cdot c_t$)			0.0423
			(0.320)
Observations	6081	6081	6081
R-Squared	0.413	0.414	0.416

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: How do retractions interact with other biases in updating (Hypothesis 2)? The sample consists in subjects in the intermediate elicitation treatment. The regressions correspond to Equation 10, using data from periods 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 $t-1$.

We first note that the second line of this equation replicates the first, except that the second line includes an interaction with the signal being in the form of a retraction. Otherwise, the first line is identical to the equation (6), which was the focus of Section 4.2 where we showed our results are consistent with the literature. Hence the presence of γ_ℓ allows us to detect how these patterns vary, depending on whether or not the signal is a retraction. In other words, the added terms provided a flexible functional form in order to capture the effect of retractions as discussed in Section 2.2, $\log(\alpha(\tau \mid s^t))$.

The results can be found in Table 3. A striking pattern emerges: while when updating from new draws we have that subjects (slightly) over-infer from signals ($\beta_2 > 1$) and do more so when signals confirm the prior ($\beta_3 > 0$), when updating from retractions they *under*-infer ($0 < \beta_2 + \gamma_2 < 1$) and exhibit *anti*-confirmation bias ($\beta_3 + \gamma_3 < 0$). In sum, belief updating from retractions 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 they the behavioral responses to retractions are not simply accentuating pre-existing biases; in fact, retractions induce opposite biases in belief reporting behavior.

While this table also demonstrates some heterogeneity in terms of when retractions are more or less likely to be effective, even these rough estimates suggest that the difference between retractions and new signals in terms of perceived informational content is practically meaningful.

4.4. 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, these observations suggest that our findings are not driven by how retractions are presented to subjects, but instead due to retractions in themselves as being treated as uniformly less informative. Lastly, we examine decision time data to understand whether it reflects the increased difficulty inherent to the contingent reasoning that interpreting retractions requires.

4.4.1. Interpreting vs. Reinterpreting Past Information

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 versus final elicitation treatments.

The results from this comparison are documented in Table 4. The specifications here correspond to equations 7 and 8 which we have described in Section 4.3.1, but including the intermediate elicitation treatment as an interaction term.[19] The result is a well-identified null result: having acted upon a piece of information has no effect on retractions.

The broad messages from this table are identical to those we have discussed in the previous sections; beliefs move in the directions of signals, but remain diminished when they are retractions relative to new signals. As far as we are able to tell, none of the 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 this hypothesis is not underlying the above-discussed distinctive manner in which individuals treat retractions.

4.4.2. The Timing of Retracted Signals

One feature of our design is that, while the signals subjects receive are exchangeable, they are observed in sequence. If subjects belief updating process displays primacy and recency effects—which have been documented in the existing literature on belief updating, see e.g. Benjamin (2019)—then it is not a priori clear whether the timing of the observed retraction should make a difference. In particular, one could conjecture that retractions would be effective in inducing 'unlearning' where they to refer to information that was just received. This Section then assesses

[19] Naturally, as when beliefs are elicited only in one period, there is no way to estimate a version of equation 9.

	(1) b$_t$	(2) b$_t$
Final (Fin_t)	0.00991 (0.0106)	0.0117 (0.0114)
Retraction (r_t)	-0.00424 (0.00343)	-0.00704 (0.00458)
Retracted Signal ($r_t \cdot s_t$)	-0.0161** (0.00717)	-0.0308*** (0.00903)
Final × Retracted Signal ($Fin_t \cdot r_t \cdot s_t$)	0.00447 (0.0109)	0.00399 (0.0170)
Final × Retraction ($Fin_t \cdot r_t$)		-0.00134 (0.00806)
Final × Signal ($Fin_t \cdot s_t$)		0.000464 (0.0101)
Compressed History FEs	Yes	No
Lagged History × Sign of Signal FEs	No	Yes
Observations	11213	9920
R-Squared	0.097	0.209

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Intermediate versus final belief elicitation. This table tests whether updating from retractions is different if beliefs have previously been elicited before a signal is retracted (Hypothesis 3). The sample is all subjects. 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 ($Fin_t \cdot r_t$ and $Fin_t \cdot s_t$ are spanned by the other controls and hence omitted, since period 3 is only in the sample when it is a retraction, making $Fin_t = Fin_t \cdot r_t$ and $Fin_t s_t = Fin_t s_t \cdot r_t$ 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. Column (1) includes fixed effects for the history at period 3. Column (2) includes fixed effects for the history at period 2 interacted with the sign of the signal.

the relevance of the timing of retracted signals in accounting for retraction failure.

The aggregated implication of timing of retracted signals is presented in Table 5. This table considers the same specifications described in Section 4.3.1, which described the diminished effectiveness of retractions, adding an indicator variable for whether the last signal observed was retracted (rf_t), as well as an interaction with the direction of the signal itself.

	(1) b$_t$	(2) b$_t$	(3) Δ b$_t$
Retraction (r_t)	0.000777	-0.00321	-0.00767
	(0.00434)	(0.00537)	(0.00468)
Retracted signal ($r_t \cdot s_t$)	-0.0289***	-0.0381***	-0.0311***
	(0.00408)	(0.00537)	(0.00468)
Last Draw Retracted (rf_t)	-0.00289	-0.00112	0.00821
	(0.00615)	(0.00714)	(0.00592)
Retracted signal x Last Draw Retracted ($r_t \cdot s_t \cdot rf_t$)	0.0156**	0.0135*	0.00581
	(0.00613)	(0.00714)	(0.00592)
Compressed history FEs	Yes	No	No
Lagged history x Sign of Signal FEs	No	Yes	Yes
Observations	17591	9074	9074
R-Squared	0.154	0.255	0.127

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Timing of the signal which is retracted. The idea is to test whether there is a difference in responding to retractions depending on whether the last signal was retracted or an earlier signal was retracted (Hypothesis 4). The sample consists in subjects in the intermediate elicitation treatment in experiment (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 a verification. We also exclude any beliefs if there was no retraction in period 3 or 4. In columns (2) and (3), we further restrict the beliefs to periods 3 and 4. In column (2), the outcome is the beliefs in period t, b_t. In column (3), the outcome is the first difference in beliefs.

We find that the absolute effectiveness of retractions is slightly increased: it is easier to disregard a piece of information if it arrived more recently, as can be seen from the fact that, in column (1), the estimated coefficient on $r_t \cdot s_t \cdot rf_t$ being strictly positive and statistically significant. However, subjects still fail to fully disregard the retracted signals, even when these were these are the most recent ones, as the mentioned coefficient is half the size of the coefficient on $r_t \cdot s_t$. As for the relative effectiveness of retractions, we find that retractions remain equally less effective than informationally-equivalent new signals in inducing changes in beliefs (columns (2) and (3)). Thus, our results suggest that, while the timing of the retracted signal may matter for the effectiveness of retractions, it does not play a major role in driving our results: a more fundamental mechanism is at play in rendering retractions less effective.

We also show disaggregated (i.e., history-by-history) versions of these results in Tables 13 through 18. The differences in the belief updates tend to be very small, even when they are significant, leading us to conclude that any differences along these dimensions are reasonably negligible; none of the comparisons are significant at the 5% level. While there are certainly differences in the average belief reports in each case, we attribute this to the fact that these comparisons have a significantly smaller number of observations. The closest we obtain to determining a difference is when the retraction is observed after two signals, which has roughly 6 times as many observations as the other comparisons. These results suggest the possibility that the retraction may be more effective when it relates to the signal that is more recent, but this difference is quite small and still not significant at the 5% level (though just barely). We conclude that in this design, timing does not have an impact; however, we emphasize that this was in part by design, since signals are observed in close succession to one another. We do not speak to whether these effects may or may not be present when information arrives over a longer timescale, and leave this to future experiments.

Similarly, we also fail to detect any particular meaning to the timing that information is *received*. In Section A.2 in the Online Appendix, we present belief reports for identical histories, which differ only on whether the retraction of a given ball color was the first or the second one observed. We fail to detect any noticeable patterns in these cases. One conclusion this observation is supportive of is that memory has a highly limited role in driving any of our results. This would be expected given our design, since subjects observe all the draws they had seen previously in every round. Still, as our design was not intended to focus on the role of memory (while this may be an interesting avenue for future work), we find it reassuring that the results do not seem to be driven by this variable. In particular, this supports or theoretical formulation, which views signals as exchangeable; as far as we can tell, this assumption seems plausible.

4.4.3. The Timing of Retractions

By having the possibility of observing either a new draw or a retraction in both periods 3 and 4, we can assess whether the timing of a retraction itself has any bearing on its effectiveness. We are motivated by the observation that beliefs are less affected by new signals the more information individuals have observed. Recalling that the effect of a signal on belief updating in this setting is constant in log-odds, existing literature has reported (log-odds) posterior beliefs to be less sensitive to new signals the more signals have been observed in the past. A possible conjecture is then that retractions ineffectiveness is conflated with this known belief updating bias. Hence, in this Section we attempt to disentangle these effects by separately considering the effect of the timing of retractions. To do so, we consider specifications similar to those described in Section 4.3.1, but comparing the effect of retractions in period 3 to those occurring 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 redefine the fixed effects for equation 8 (column (2)) on $H_{t-2} \times \{s_{t-1}\} \times \{s_t\}$. This implies we are restricting attention to comparing the cases where the compressed history at period 4 is the same but in one the retraction took place in period 3 and in the other in period 4 and both retracted the same signal. That is, we are 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 $\tau \in \{1, 2\}$. We do not estimate equation 9 as we cannot both control for the histories as desired and compare $(s_1, s_2, s_3, n_\tau = 1)$ to the case of obtaining an informationally-equivalent new signal $(s_1, s_2, s_3, s_4 = -s_\tau)$.

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.

4.4.4. Retractions are Harder to Process

We then turn to another mechanism which may be behind retraction failure: that retractions are simply harder to process. If retractions as information about information are noticeably harder to process, a natural proxy for processing difficulty is the subjects' decision time d_t, under the assumption that the greater the difficulty the longer the time taken to interpret the information being provided.

To test this hypothesis we regress decision time on a dummy variable indicating whether or not a retraction occurs in that period. As before, we control for the lagged history interacted with the direction of the signal or retraction observed, such that we compare decision times for retraction to those of informationally equivalent new signals. We estimate

$$d_t = \beta_0 + \beta_1 \cdot r_t + F_{H_{t-1} \times \{s_t\}}, \tag{11}$$

as well as a variant of this equation using log decision time as the independent variable instead.

	(1)	(2)
	b_t	b_t
Retraction ($r_3 + r_4$)	0.00773	0.00737
	(0.00630)	(0.00850)
Retraction in Period 4 (r_4)	-0.00104	-0.00194
	(0.00836)	(0.00838)
Retracted Signal ($r_3 \cdot s_3 + r_4 \cdot s_4$)	-0.0368***	-0.0417***
	(0.00596)	(0.00742)
Retracted Signal × Retraction in Period 4 ($r_4 \cdot s_4$)	0.00218	0.0154
	(0.00836)	(0.00982)
Compressed History FEs	Yes	No
Lagged Lagged History × Lagged Sign of Signal × Sign of Signal FEs	No	Yes
Observations	9432	4350
R-Squared	0.203	0.274

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Timing of retraction: does the timing at which a retraction is received matter for its effect (Hypothesis 5)? 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 fixed up to order. The sample consists in subjects in the intermediate elicitation 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 comparison is, for beliefs in period 4, 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.

	(1)	(2)
	d_t	$\log(\mathrm{d}_t)$
Retraction (r_t)	0.493***	0.100***
	(0.0908)	(0.0128)
Lagged history * Sign of Signal FEs	Yes	Yes
Mean of dep. variable	5.57	1.57
Observations	8986	8986
R-Squared	0.010	0.016

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Decision times. This table tests whether the time taken to report beliefs is different after retractions compared to equivalent new signals (Hypothesis 6). The sample consists in subjects in the intermediate elicitation 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).

The results, which can be seen in Table 7, confirm our conjecture: subjects take longer to report their beliefs when updating from retractions. Thus, the data suggests that retractions are not only treated differently, they are harder to process.

We conjecture that this increased complexity is due to the need of a particular kind of contingent reasoning which is inherent to interpreting information about information. As discussed earlier, other work has identified failures in belief updating specifically due to failures of contingent reasoning. But, even in a general model of belief updating that subsumes many specific instances of non-Bayesian updating, retractions are treated differently from new signals if and only if a particular instance of contingent reasoning fails, entailing $\log\left(\alpha(\tau \mid s^t)\right) \neq 0$ (Proposition 1). We interpret our results in Table 7 as providing supporting evidence that the contingent reasoning involved in updating from retractions is distinct from and more complex than that involved in updating from new signals.

4.5. Updating After Retractions

Our design also allowed us to see how subjects would respond to belief updates *following* the retraction of information. While a number of studies have investigated the effect of retractions in the most varied settings, we are unaware of work examining how these affect the interpretation of ensuing information. Were individuals to discount further evidence following a retraction, then retractions are not only less effective than new direct evidence, they would also hamper the interpretation of future evidence, further dampening the absorption of new information. Our results speaking to this can be found in Table 8. We test whether observing a retraction affects *subsequent* updating in two ways. First, we maximize our use of the experimental variation, plus fixed effects, in columns (1)-(4) of Table 8, by estimating two variants of the following equation:

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

In columns (1) and (3), 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 (2) and (4), we restrict the sample to period 4 and we have fixed effects F corresponding to the history at period 2 interacted with the sign of the signal in period 3, i.e. $F_{H_{t-2}\times\{s_{t-1}\}}$. As such, we can compare the change in belief

	(1) Δ b_t	(2) Δ b_t	(3) Δ b_t	(4) Δ b_t
Signal (s_t)	0.0574***	0.0579***	0.0574***	0.0578***
	(0.00172)	(0.00343)	(0.00172)	(0.00343)
Retraction in Previous Period (r_{t-1})	0.00826**	0.00950*	0.00816**	0.00989*
	(0.00395)	(0.00546)	(0.00395)	(0.00546)
Signal × Retraction in Previous Period ((s_t) $\cdot$ (r_{t-1}))	0.00900**	0.0154***	0.00901**	0.0157***
	(0.00389)	(0.00513)	(0.00389)	(0.00513)
Retracted Signal in Previous Period ($r_{t-1} \cdot s_{t-1}$)			0.00127	0.0131**
			(0.00339)	(0.00546)
Signal × Retracted Signal in Previous Period ((s_t) $\cdot$ ($r_{t-1} \cdot s_{t-1}$))			0.00386	0.00249
			(0.00339)	(0.00391)
Lagged Compressed History FEs	Yes	No	Yes	No
Lagged Lagged History × Lagged Sign of Signal FEs	No	Yes	No	Yes
Observations	12209	5457	12209	5457
R-Squared	0.112	0.103	0.112	0.104

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Signals after retractions. This tests updating from new signals after retractions, compared to after the equivalent compressed history, and also compared to after the equivalent new signal (Hypothesis 7). The sample consists in subjects in the intermediate elicitation 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.

	(1) d_t	(2) log(d_t)
Lagged Retraction (r_{t-1})	0.224*	0.0441**
	(0.130)	(0.0179)
Lagged Lagged History × Lagged Sign of Signal × Sign of Signal FEs	Yes	Yes
Mean of dep. variable	5.93	1.61
Observations	5405	5405
R-Squared	0.010	0.014

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Decision times after retractions. This table tests whether the time taken to report beliefs is different after retractions compared to equivalent new signals. The sample consists in subjects in the intermediate elicitation 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.

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.

We conclude our analysis by checking whether experiencing a retraction in the past results in an increased level of effort or attention by the subjects when considering new signals. Similarly to before, we estimate the effect of a retraction on how long subjects take in subsequent updating from new signals by estimating

$$d_t = \beta_0 + \beta_1 \cdot r_{t-1} + F_{H_{t-2} \times \{s_{t-1}\} \times \{s_t\}}, \tag{12}$$

where we control for twice-lagged history and the signs of the signals in the previous and in the current periods. Analogously to Section 4.4.4, we also report on the effect on 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 and statistically significant only when considering log decision time.

5. 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 well-known biases. Our analysis suggests that retraction failure is due to a failure of contingent reasoning particular to information about information—rendering retractions harder to interpret—and not just an expression of these well-known biases. Table 10 revisits each of these hypotheses, and assesses our findings.

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. In particular, we see two natural directions for follow-on work given these observations.

First, exploring this phenomenon in particular contexts, where it may interact with other

Hypothesis	Documented (✓) or not detected (✗)	Other Comments
1, Part (a): Subjects fail to fully internalize retractions	✓	
1, Part (b): Subjects treat retractions as less informative than equivalent new information	✓	
2: Updating from retractions accentuates biases in updating	✓	Retractions reverse the direction of the biases to under-inference and anti-confirmation bias
3: Retractions have less of an effect when subjects have acted upon observed signals	✗	Precise null
4: Retractions influence beliefs differently depending on when the retracted signal was observed.	✗	Imprecise null; possible that retractions are slightly more effective for 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	✓	More sensitivity to signals after retractions

Table 10: Summary of our assessment of our 7 main hypotheses. See Section 2.3 for a more complete description of each, as well as the reasoning involved with formulating each one.

behavioral biases, is likely to generate further insights. 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? As mentioned in our review of the literature, a significant body of work on political behavior suggests that many factors are at play which interfere with Bayesian reasoning. Any such nuances, we believe, could yield insights which help understand the failure or success of retractions in practice, and of fact-checking more generally.

Second, we have not fully explored the possible heterogeneity in the reactions to retractions. In particular, 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. 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.

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

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

Arechar, A., S. Gächter, and L. Molleman (2018): "Conducting interactive experiments online," *Experimental Economics*, 21, 99–131.

Benabou, R. and J. Tirole (2016): "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 30, 141–164.

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.

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.

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

Brunnermeier, M. and J. Parker (2005): "Optimal Expectations," *American Economic Review*, 144, 1092–1118.

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

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.

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.

Conlon, J. J., M. Mani, G. Rao, M. Ridley, and F. Schilbach (2021): "Learning in the Household," *Working Paper*.

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

——— (2019b): "Testing Models of Belief Bias: An Experiment," *Games and Economic Behavior*, 133, 549–565.

Cripps, M. (2019): "Divisible Updating," *Working Paper*.

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

Eil, D. and J. Rao (2011): "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *AEJ: Microeconomics*, 3, 114–138.

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

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

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

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

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

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

Grossman, Z. and D. Owens (2012): "An Unlucky Feeling: Overconfidence and Noisy Feedback," *Journal of Economic Behavior and Organization*, 84, 510–524.

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

Huber, G., S. Hill, and G. Lenz (2012): "Sources of Bias in Retrospective Decision Making: Experimental Evidence on Voters' Limitations in Controlling," *American Political Science Review*, 106, 720–741.

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

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.

Kovach, M. (2020): “Twisting the truth: Foundations of wishful thinking,” *Theoretical Economics*, 15, 989–1022.

Landier, A., Y. Ma, and D. Thesmar (2020): “Biases in Expectations: Experimental Evidence,” *Working Paper.*

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.

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

Martínez-Marquina, A., M. Niederle, and E. Vespa (2019): “Failures in Contingent Reasoning: The Role of Uncertainty,” *American Economic Review*, 109, 3437–3474.

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.

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

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

Oprea, R. and S. Yuksel (2020): “Social Exchange of Motivated Beliefs,” *Working Paper.*

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

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

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

Snowberg, E. and L. Yariv (2020): “Testing the Waters: Behavior across Participant Pools,” *American Economic Review*, Forthcoming.

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

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

Online Appendix for Learning versus Unlearning

A. TABLES AND FIGURES

A.1. Basic Data on Beliefs

Figure 4: Distribution of reported beliefs

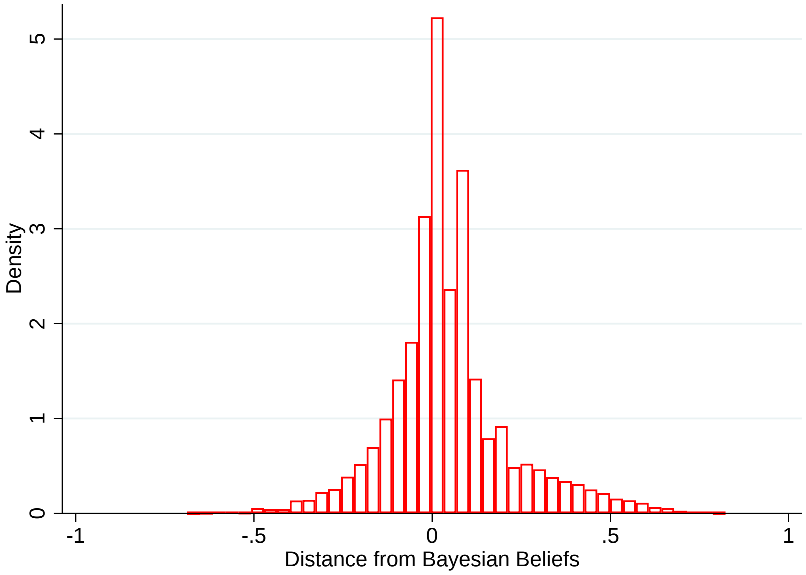
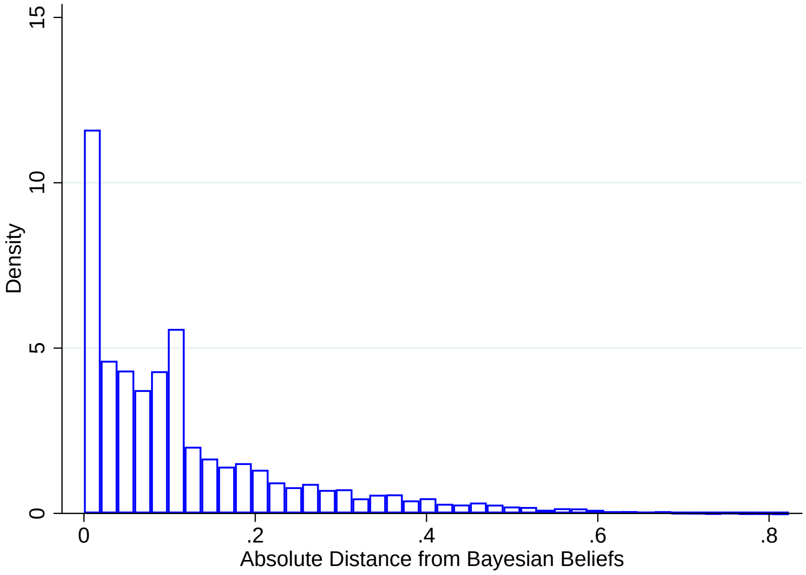


Figure 5: Distribution of reported beliefs



	(1)	(2)
Bayesian Beliefs	0.614^{***}	0.745^{***}
	(0.0163)	(0.0163)
Constant	0.187^{***}	0.129^{***}
	(0.00956)	(0.00968)
N	17896	15521

Standard errors in parentheses

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

Table 11: Sanity check. Regressing stated beliefs against their Bayesian benchmark. This compares belief reports to Bayesian beliefs. In the second specification, we exclude participants who updated in the wrong direction.

	(1)	(2)	(3)	(4)	(5)
	l$_t$	l$_t$	l$_t$	l$_t$	l$_t$
Prior (l_{t-1})	0.883***		0.716***	0.883***	0.938***
	(0.0233)		(0.0257)	(0.0371)	(0.0353)
Signal (s_t)	1.321***	0.621***	0.709***	1.417***	1.566***
	(0.0463)	(0.0331)	(0.0259)	(0.0628)	(0.117)
Observations	11739	6752	6752	3317	1670
R-Squared	0.425	0.049	0.498	0.385	0.507

Standard errors in parentheses

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

Table 12: Updating from new draws – by period. This table represents updating from standard new ball draws across different periods. The sample consists in subjects in the intermediate elicitation 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. Thus, for example, if there is a retraction in period 3, we exclude beliefs in both period 3 and 4. Column (1) uses data of all periods. Columns (2) through 5 use a sample restricted to periods 1 through 4, respectively. The regressions correspond to Equations 4, 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 $t-1$. .

A.2. Comparisons on Time of the Retracted Signal

	(1)
VARIABLES	Beliefs
Case 1: 11, Retracting signal from period 1	0.00336
	(0.00788)
Case 2: 11, Retracting signal from period 2	-0.0162**
	(0.00775)
Constant	0.576***
	(0.00321)
Observations	4,495
R-squared	0.001
Retracting signal from period 1 = Retracting signal from period 2	0.0522

Table 13: Case 1.1: 11. Comparing belief updating with retraction of the first signal vs. the second signal.

VARIABLES	(1) Beliefs
Case 1: 111, Retracting signal from period 1	0.00555
	(0.0288)
Case 2: 111, Retracting signal from period 2	0.0482*
	(0.0259)
Constant	0.585***
	(0.00901)
Observations	777
R-squared	0.005
Retracting signal from period 1 = Retracting signal from period 2	0.244

Table 14: Case 1.2: 111. Comparing belief updating with retraction of the first signal vs. the second signal.

VARIABLES	(1) Beliefs
Case 1: 110, Retracting signal from period 1	-0.0600***
	(0.0121)
Case 2: 110, Retracting signal from period 2	-0.0679***
	(0.0123)
Constant	0.588***
	(0.00768)
Observations	1,339
R-squared	0.027
Retracting signal from period 1 = Retracting signal from period 2	0.556

Table 15: Case 1.3: 110. Comparing belief updating with retraction of the first signal vs. the second signal.

	(1)
VARIABLES	Beliefs
Case 2: 111, Retracting signal from period 1	0.0163
	(0.0288)
Case 3: 111, Retracting signal from period 3	0.0438*
	(0.0259)
Constant	0.574***
	(0.00895)
Observations	824
R-squared	0.004
Retracting signal from period 1 = Retracting signal from period 3	0.453

Table 16: Case 2.1: 111. Comparing belief updating with retraction of the first signal vs. the third signal.

	(1)
VARIABLES	Beliefs
Case 1: 101, Retracting signal from period 1	-0.00405
	(0.0102)
Case 3: 101, Retracting signal from period 3	-0.00362
	(0.0158)
Constant	0.533***
	(0.00681)
Observations	945
R-squared	0.000
Retracting signal from period 1 = Retracting signal from period 3	0.979

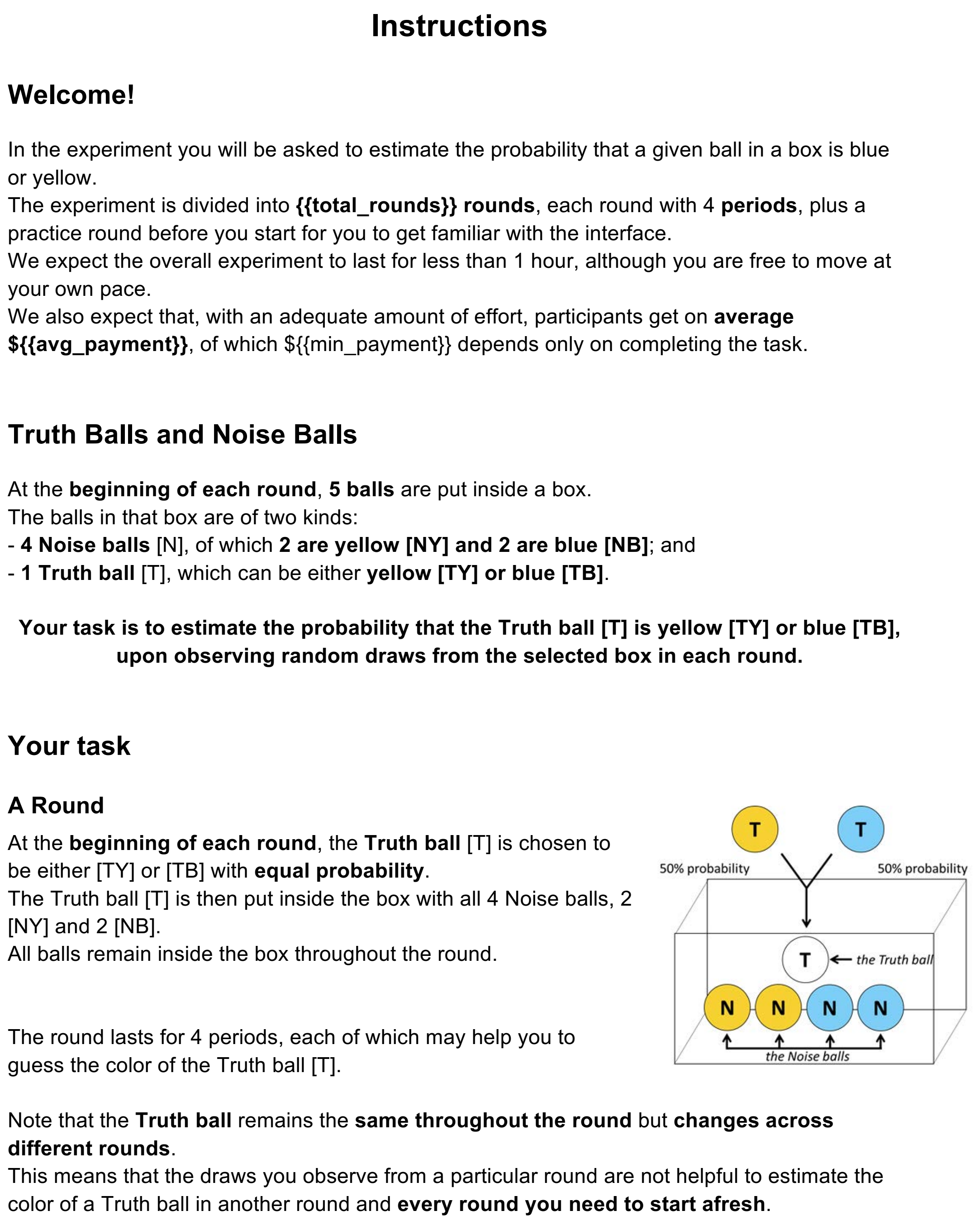
Table 17: Case 2.2: 101. Comparing belief updating with retraction of the first signal vs. the third signal.

VARIABLES	(1) Beliefs
Case 2: 100, Retracting signal from period 2	0.0113
	(0.0116)
Case 3: 100, Retracting signal from period 3	0.00112
	(0.0215)
Constant	0.467***
	(0.00879)
Observations	933
R-squared	0.001
Retracting signal from period 2 = Retracting signal from period 3	0.627

Table 18: Case 3.1: 100. Comparing belief updating with retraction of the second signal vs. the third signal.

B. INSTRUCTIONS

B.1. Instructions (June 2020), not including preamble



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 **{{total_rounds}} rounds**, each round with 4 **periods**, plus a practice round 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 ${{avg_payment}}**, of which ${{min_payment}} 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** [N], of which **2 are yellow [NY] and 2 are blue [NB]**; and
- **1 Truth ball** [T], which can be either **yellow [TY] or blue [TB]**.

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

Your task

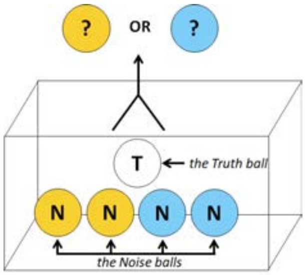
A Round

At the **beginning of each round**, the **Truth ball** [T] is chosen to be either [TY] or [TB] with **equal probability**.
The Truth ball [T] is then put inside the box with all 4 Noise balls, 2 [NY] and 2 [NB].
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 [T].

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**.

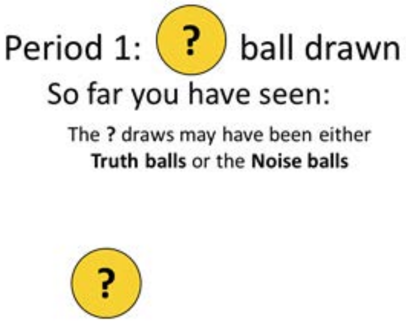
Periods 1 and 2



In periods 1 and 2, **a ball is drawn** from the box **at random** and you are told its **color**, [Y] or [B].
The **ball** is then **placed back** into the box.
You will not be told whether it is a Noise ball [N] or the Truth ball [T]. 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 [T] with 20% probability**;
- is a **Noise ball [N] 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.

This is an example of what you can see at period 1:



Period 1: ? ball drawn

So far you have seen:

The **?** draws may have been either **Truth balls** or the **Noise balls**

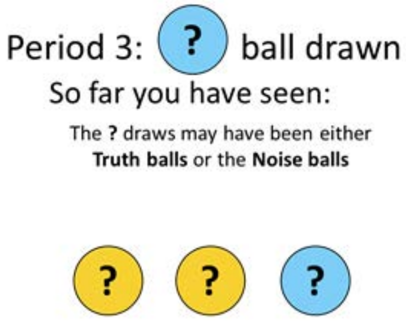
Periods 3 and 4

At the beginning of periods 3 and 4, a coin is flipped, and
(i) with 50% probability it lands heads and you will observe a **new draw** from the box,
(ii) with 50% probability it lands tails and you will observe a **retraction**, learning whether one of the balls is a Noise ball or the Truth ball.

(i) New Draw

If you get a **new draw**, it will be **exactly as before**: a ball is drawn from the box and its color is shown to you, but not whether it is the Truth ball or the Noise ball.
Since the balls are drawn at random, the drawn ball [?]:
- is the Truth ball [T] with 20% probability
- is a Noise ball [N] with 80% probability.

This is an example of what you can see if you get a new draw in period 3:



Period 3: ? ball drawn

So far you have seen:

The **?** draws may have been either
Truth balls or the **Noise balls**

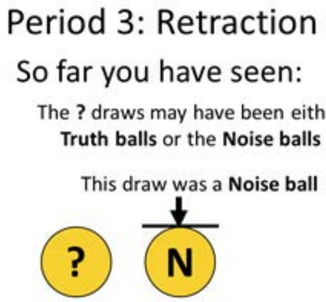
(ii) Retraction

If you get a retraction,

one of the [?] draws is chosen at random with equal probability, regardless of whether they were draws of the Truth [T] or Noise [N] balls.

You are then **showed whether** that **draw** was a **Noise ball [N]** or the **Truth ball [T]** itself.

This is an example of what you can see if you get a retraction in period 3:



Period 3: Retraction

So far you have seen:

The **?** draws may have been either
Truth balls or the **Noise balls**

This draw was a **Noise ball**

New Round

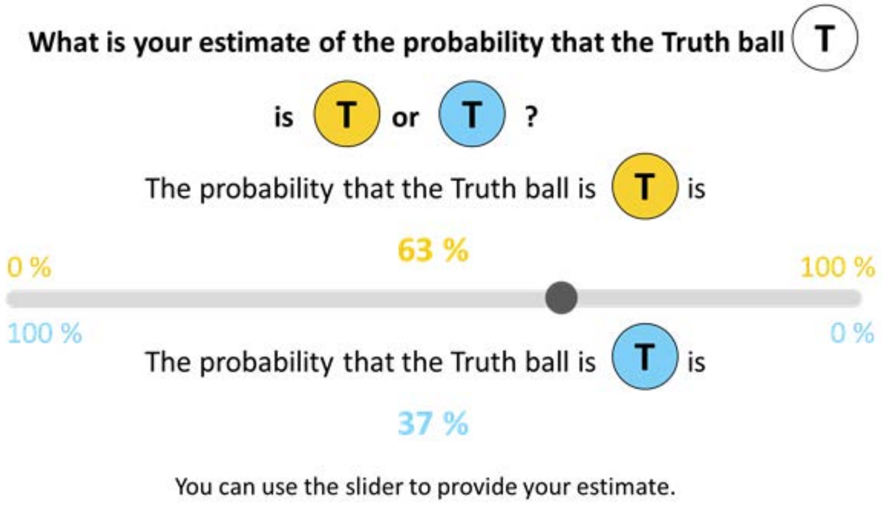
After these 4 periods, a new round begins.
Each round, a new color for the Truth ball [T] is selected the same way and independently.
This means that **whether the Truth ball is [TY] or [TB] in one round has no influence on whether the Truth ball is [TY] or [TB] in another round**.
It will be clearly indicated when a new round begins.

Estimates

Every period and every round you will be asked to provide your estimate of the probability that the Truth ball [T] is yellow [TY] or blue [TB].

Unless it is shown to you in a retraction, you will not be able to know the color of the Truth ball for sure, but you will be able to make **inferences based on the draws** you have seen.
You will be **paid based on** how **accurate** your estimate is.

You can enter your estimate using the slider.



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

63 %

The probability that the Truth ball is T is

37 %

You can use the slider to provide your estimate.

Payment

By completing the experiment, you can secure ${{min_payment}} for sure.

You can get a bonus of an additional ${{bonus_payment}} depending on your performance.

At each period, you will receive a number of points which depends on your estimate and on the color of the Truth ball [T] in that round.

The higher the probability you assign to the correct color,
the more points you get at each round.

If your estimate in a given period is that the Truth ball is [TY] with probability q (x 100%) and an [TB] with probability 1-q (x 100%), then you will receive
100 x (1 - (1-q)^2) points if the Truth ball is [TY], and
100 x (1 - q^2) points if the Truth ball is [TB].

So if your estimate completely correctly the color of the Truth ball, you get 100 points and if you estimate completely incorrectly you get 0 points.

The fewer probability you assign to the correct color,
the fewer points you receive.

For instance, if you estimate that the Truth ball is [TY] with 89% probability and [TB] with 11% probability, you receive 98.79 points if the Truth ball is indeed [TY] and 20.79 if the Truth ball is instead [TB].

The points you get determine the probability of you getting the bonus.

In order to determine the probability of you getting the bonus, at the end of the experiment, one of the rounds is picked randomly with equal probability and, in this round, one of the periods is then chosen randomly, with equal probability.

The **points you got = probability of getting the $6 bonus**.
This means that if in the selected round/period you have 99.84 points you have 99.84% probability of getting the $6 bonus. If you have 36 points you only have 36% probability.

There is, of course, an element of chance in the task, but **the more you pay attention, the more you increase the probability of getting the bonus**.

All in all, the implication of the reward rule is straightforward: To maximize your expected earnings, the **best** thing you can do in each period is to always **report your best estimate** of the probability that the Truth ball is [TY] or [TB].

This reward system has been designed to encourage you to provide your best estimates.

Questionnaire

After you have completed all rounds, we will ask you some quantitative reasoning questions, for which you can get an extra ${{quant_bonus}} in bonus and then generic demographic questions.
We will **not** be collecting any information that allows us to identify you.
The data will be anonymized and your MTurk ID will **not** be available.
This data will be used for *scientific research purposes only*.

Only after you answer these questions will the task be completed and we will proceed to implement payments.

Questions

Q1:
How many Noise Balls are there?
0, 1, 2, 3, [✓ 4]

Q2:
How many of the Noise Balls are [NY] and [NB]?
1 [NY] and 3 [NB]
3 [NY] and 1 [NB]
[✓ 2 [NY] and 2 [NB]]

Q3:
It is possible that you see a [?Y] ball 4 times and the Truth ball is [TB].
[✓ The statement is true.]
The statement is false.

Q4:
Even if in a given round the Truth Ball is [TY], in the following round the Truth ball can either be [TY] or [TB] with equal (50 % -- 50 %) probability.
[✓ The statement is true.]
The statement is false.

Q5:
If a draw you were shown [?Y] corresponded to a Noise ball [NY], then it means the Truth ball has to be [TB] and not [TY].
The statement is true.
[✓ The statement is false.]

Q6:

If a draw you were shown [?Y] corresponded to a Noise ball [NY], then it means the Truth ball [T] may or may not be of a different color.

[✓ The statement is true.]

The statement is false.

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 **{{total_rounds}} rounds**, each round with either 2 or 3 **periods**, plus a practice round 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 ${{avg_payment}}**, of which ${{min_payment}} depends only on completing the task.

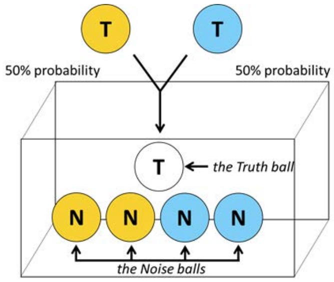
Truth Balls and Noise Balls

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** [N], of which **2 are yellow [NY] and 2 are blue [NB]**; and
- **1 Truth ball** [T], which can be either **yellow [TY] or blue [TB]**.

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

Your task

A Round

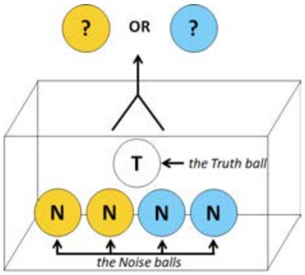


At the **beginning of each round**, the **Truth ball** [T] is chosen to be either [TY] or [TB] with **equal probability**.
The Truth ball [T] is then put inside the box with all 4 Noise balls, 2 [NY] and 2 [NB].
All balls remain inside the box throughout the round.

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

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**.

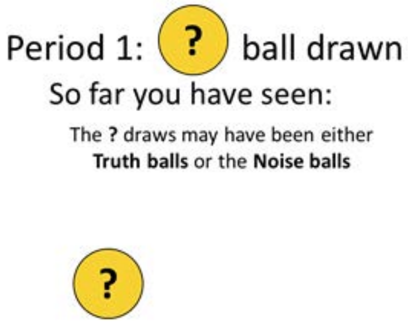
Periods 1 and 2



In periods 1 and 2, **a ball is drawn** from the box **at random** and you are told its **color**, [Y] or [B].
The **ball** is then **placed back** into the box.
You will not be told whether it is a Noise ball [N] or the Truth ball [T]. 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 [T] with 20% probability**;
- is a **Noise ball [N] 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.

This is an example of what you can see at period 1:



Period 1: ? ball drawn

So far you have seen:

The **?** draws may have been either
Truth balls or the **Noise balls**

At the end of period 2, with 33% probability the round ends, and with 66% probability you will move to period 3.

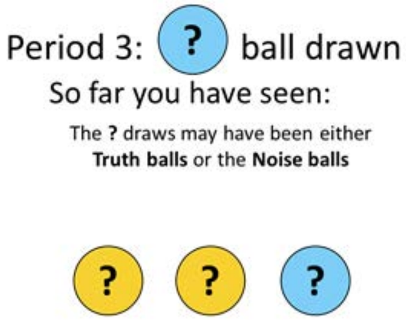
Period 3

At the beginning of period 3, a coin is flipped, and
(i) with 50% probability it lands heads and you will observe a **new draw** from the box,
(ii) with 50% probability it lands tails and you will observe a **retraction**, learning whether one of the balls is a Noise ball or the Truth ball.

(i) New Draw

If you get a **new draw**, it will be **exactly as before**: a ball is drawn from the box and its color is shown to you, but not whether it is the Truth ball or the Noise ball.
Since the balls are drawn at random, the drawn ball [?]:
- is the Truth ball [T] with 20% probability
- is a Noise ball [N] with 80% probability.

This is an example of what you can see if you get a new draw in period 3:



Period 3: ? ball drawn

So far you have seen:

The **?** draws may have been either **Truth balls** or the **Noise balls**

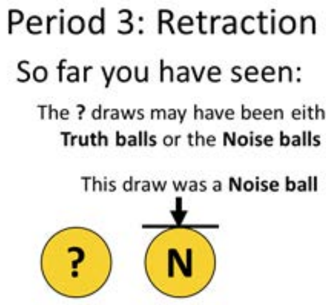
(ii) Retraction

If you get a retraction,

one of the [?] draws is chosen at random with equal probability, regardless of whether they were draws of the Truth [T] or Noise [N] balls.

You are then **showed whether** that **draw** was a **Noise ball [N]** or the **Truth ball [T]** itself.

This is an example of what you can see if you get a retraction in period 3:



Period 3: Retraction

So far you have seen:

The **?** draws may have been either **Truth balls** or the **Noise balls**

This draw was a **Noise ball**

New Round

The round can end after period 3, with 2/3 probability (66.7%), or after period 2, with 1/3 probability (33.3%), in which case you will skip period 3.

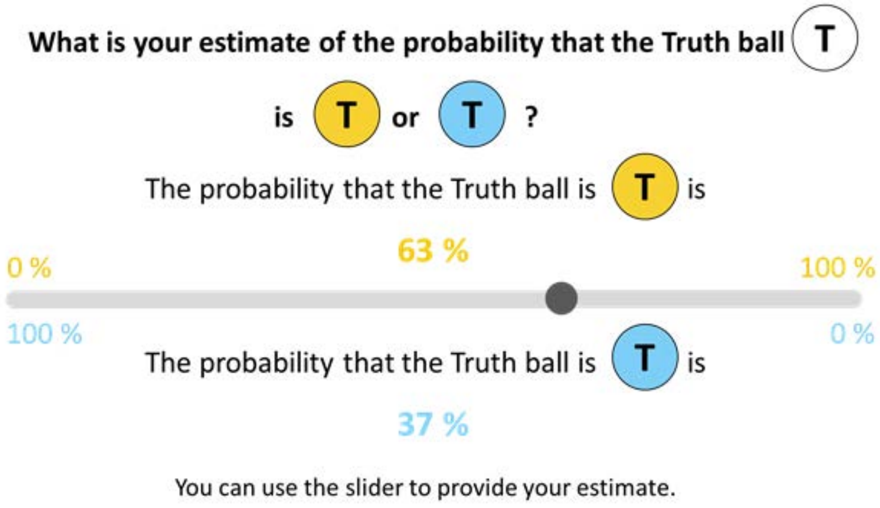
After the round ends, a new round begins.
Each round, a new color for the Truth ball [T] is selected the same way and independently.
This means that **whether the Truth ball is [TY] or [TB] in one round has no influence on whether the Truth ball is [TY] or [TB] in another round**.
It will be clearly indicated when a new round begins.

Estimates

At the end of every round you will be asked to provide your estimate of the probability that the Truth ball [T] is yellow [TY] or blue [TB].

Unless it is shown to you in a retraction, you will not be able to know the color of the Truth ball for sure, but you will be able to make **inferences based on the draws** you have seen.
You will be **paid based on** how **accurate** your estimate is.

You can enter your estimate using the slider.



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

63 %

The probability that the Truth ball is T is

37 %

You can use the slider to provide your estimate.

Payment

By completing the experiment, you can secure ${{min_payment}} for sure.
You can get a bonus of an additional ${{bonus_payment}} depending on your performance.

At each round, you will receive a number of points which depends on your estimate and on the color of the Truth ball [T] in that round.

The higher the probability you assign to the correct color,
the more points you get at each round.

If your estimate in a given round is that the Truth ball is [TY] with probability q (x 100%) and an [TB] with probability 1-q (x 100%), then you will receive
100 x (1 - (1-q)^2) points if the Truth ball is [TY], and
100 x (1 - q^2) points if the Truth ball is [TB].

So if your estimate completely correctly the color of the Truth ball, you get 100 points and if you estimate completely incorrectly you get 0 points.

The fewer probability you assign to the correct color,
the fewer points you receive.

For instance, if you estimate that the Truth ball is [TY] with 89% probability and [TB] with 11% probability, you receive 98.79 points if the Truth ball is indeed [TY] and 20.79 if the Truth ball is instead [TB].

The points you get determine the probability of you getting the bonus.

In order to determine the probability of you getting the bonus, at the end of the experiment, one of the rounds is picked randomly with equal probability.

The **points you got = probability of getting the $6 bonus**.
This means that if in the selected round you have 99.84 points you have 99.84% probability of getting the $6 bonus. If you have 36 points you only have 36% probability.

There is, of course, an element of chance in the task, but **the more you pay attention, the more you increase the probability of getting the bonus**.

All in all, the implication of the reward rule is straightforward: To maximize your expected earnings, the **best** thing you can do in each round is to always **report your best estimate** of the probability that the Truth ball is [TY] or [TB].

This reward system has been designed to encourage you to provide your best estimates.

Questionnaire

After you have completed all rounds, we will ask you some quantitative reasoning questions, for which you can get an extra ${{quant_bonus}} in bonus and then generic demographic questions.
We will **not** be collecting any information that allows us to identify you.
The data will be anonymized and your MTurk ID will **not** be available.
This data will be used for *scientific research purposes only*.

Only after you answer these questions will the task be completed and we will proceed to implement payments.

Questions

Q1:
How many Noise Balls are there?
0, 1, 2, 3, [✓ 4]

Q2:
How many of the Noise Balls are [NY] and [NB]?
1 [NY] and 3 [NB]
3 [NY] and 1 [NB]
[✓ 2 [NY] and 2 [NB]]

Q3:
It is possible that you see a [?Y] ball 3 times and the Truth ball is [TB].
[✓ The statement is true.]
The statement is false.

Q4:
Even if in a given round the Truth Ball is [TY], in the following round the Truth ball can either be [TY] or [TB] with equal (50 % -- 50 %) probability.
[✓ The statement is true.]
The statement is false.

Q5:
If a draw you were shown [?Y] corresponded to a Noise ball [NY], then it means the Truth ball has to be [TB] and not [TY].
The statement is true.
[✓ The statement is false.]

Q6:

If a draw you were shown [?Y] corresponded to a Noise ball [NY], then it means the Truth ball [T] may or may not be of a different color.

[✓ The statement is true.]

The statement is false.

Instructions: Quantitative Questions

Below you can see 3 different questions.
For each question, choose the option that you think is correct.
There is only one correct answer for each question.

One of these 3 questions will be chosen randomly with equal probability.
If your answer to the chosen question is correct, you will get an addition ${{quant_bonus}}.
If your answer is not correct, you get no additional money.

Q1

Q2

Q3