Stories, Statistics, and Memory*

Thomas Graeber Christopher Roth Florian Zimmermann

October 2, 2023

Abstract

For most decisions, we encounter relevant information over the course of days, months or years. We consume such information in various forms, including quantitative data about collections of observations – statistics – and qualitative content about individual instances – stories. This paper proposes that the information type – story versus statistic – shapes selective memory. In controlled experiments, we document a pronounced story-statistic gap in memory: the average impact of stories on beliefs fades by 33% over the course of a day, but by 73% for statistics. Guided by a model of selective memory, we disentangle different mechanisms underlying the story-statistic gap. The similarity between relevant information in memory and the prompt drives the gap. Irrelevant information that is similar to the prompt, on the other hand, impedes successful recall.

Keywords: Memory; Belief Formation; Stories; Narratives; Statistical Information.

*We thank the editors and five anonymous referees for very helpful and constructive suggestions. We also thank seminar audiences at the Belief-Based Utility Workshop in Amsterdam, UC Berkeley, the Berlin Behavioral Economics Seminar, Bocconi, the Booth School of Business, the briq Beliefs Workshop, CERGE-EI Prague, the CESifo Conference on Behavioral Economics, Cologne, Cornell, Harvard, Hebrew University, Heidelberg, Innsbruck, MiddExLab, Sciences Po, the Spring School of Behavioral Economics, Stanford, and UCLA Anderson for helpful comments. We thank Andrea Amelio, Simon Cordes, Paul Grass, Tsahi Halyo, Apoorv Kanoongo, Emir Kavukcu, Malte Kornemann, Robin Musolff, Constantin Schesch, and Malin Siemers for excellent research assistance. Funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2126/1-390838866 is acknowledged. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant: 948424 - MEMEB). The research described in this article was approved by the Institutional Review Board at the University of Cologne. Graeber: Harvard Business School, tgraeber@hbs.edu. Roth: University of Cologne, ECONtribute, CEPR, briq and MPI for Collective Goods Bonn, roth@wiso.uni-koeln.de. Zimmermann: University of Bonn and briq, florian.zimmermann@briq-institute.org.

1 Introduction

On many economic, political, and cultural issues, we accumulate a wealth of relevant information over the course of time. However, when making a decision, we may often remember only a subset of that information. This paper studies the nature of such imperfect recall and its influence on the evolution of beliefs. In practice, the type of information we are exposed to ranges from collections of information shown in numbers[1] – statistics – to qualitative descriptions of a single or few instances – stories. In this paper, we examine the hypothesis that we are more likely to successfully retrieve stories than statistics. If stories, even when unrepresentative of reality at large, are more easily recalled than statistics, people's beliefs may be disproportionately influenced by information they received in the form of stories. As time passes, this selective recall can give rise to misperceptions about reality, and bears implications for how policymakers and managers should communicate to persistently persuade their audiences.

To study the temporal evolution of beliefs in response to different types of information, we run a series of tightly controlled, pre-registered experiments. In our baseline study, a hypothetical product or venue has received a number of reviews, each either positive or negative. We induce a uniform prior over the number of positive reviews. We study different types of information: quantitative information about the number of positive reviews in a randomly drawn subsample of reviews (the *Statistic* condition), information about a randomly drawn single review alongside qualitative content describing the experience underlying the review (the *Story* condition), and no additional information. Participants are asked to guess whether another randomly drawn review is positive. In a within-subject design, each participant is presented with three independent scenarios, and the type of additional information they receive is randomized for each scenario. To examine the role of memory, we elicit beliefs from participants twice: immediately (the *Immediate* condition) and again following a one-day delay (the *Delay* condition). The temporal structure is crucial to our study design as there are numerous differences between stories and statistics that could result in different beliefs; however, any such differences *not* associated with memory are accounted for by the immediate belief update. Therefore, since no new information is received in the interim, any change in stated beliefs over time must, by design, be due to memory.

We document that, in line with imperfect memory, average beliefs partially revert to the prior as time passes for both types of information. Our main finding is a pronounced story-statistic gap in the evolution of beliefs: the effect of stories on beliefs decays less strongly than the effect of statistics. Pooling all statistics and stories presented in our baseline study, we find that, on average, the magnitude of belief reversion from the

[1]This is the Oxford dictionary's definition of a statistic.

immediate update towards the prior is more than twice as large for statistics (73%) as for stories (33%). In fact, we find that the average belief impact (the difference between a stated immediate or delayed belief and the induced prior) is larger for statistics than stories in *Immediate*, but smaller in *Delay*. This means that the relative magnitudes of belief impact *flip* over time: statistics are, on average, interpreted as more informative in the moment and shift beliefs more strongly, but as time passes this effect reverses and the effect of stories on beliefs is larger. To provide direct evidence on the role of selective memory, we use a free recall task in the follow-up survey. We find that participants are more accurate at recalling the correct type and direction of the information for the scenario in which they received a story than for the one in which they received a statistic.

Additional experiments demonstrate the robustness of our findings in a setting where respondents receive stories with qualitative content that is known to be uninformative, which allows us to compare belief movements relative to Bayesian benchmarks for both stories and statistics. Our experiments replicate our previous findings and establish that immediate belief updates are fairly close to Bayesian benchmarks for both types of information. Moreover, we conduct an experiment which shows that adding uninformative qualitative content to statistical information also significantly decreases belief decay and increases accurate recall. This suggests that it is the qualitative features of stories that drive the story-statistic gap.

To guide our investigation of underlying memory mechanisms, we propose a simple model of selective memory that adapts Bordalo et al. (2023a,c) to accommodate stories and statistics. Our framework follows canonical models of memory structure in the literature, where experiences are stored as memory traces in episodic memory. A central feature of episodic memory is its cue-dependent nature: the recall of memory traces is triggered based on their relationship to a cue. Here, the memory cue is generated by a prompt to state a belief or recall information previously received in a given scenario. The scenario prompt evokes thoughts that are related in meaning to it, and these *semantic associations* with the scenario constitute the cue. For example, the scenario prompt "restaurant" is connected in meaning to concepts such as dinner, food and drinks (among many others), all of which would form part of the cue. Given the cue, recall is stochastic: the decision-maker either recalls the relevant target memory associated with the prompted scenario or they accidentally recall an irrelevant, non-target memory that does not originate from the scenario. Whether a relevant or an irrelevant memory is retrieved is governed by two similarity relationships. The more similar a target memory trace is to the cue, i.e. the higher the *cue-target similarity*, the higher the chances of retrieving it. The more similar non-target memory traces are to the cue, the higher the chances of failing to retrieve the target and accidentally retrieving an irrelevant memory. This mechanism is referred to as *interference* in the memory process.

Our simple framework makes three testable predictions about the likelihood of successful recall and the temporal decay of the initial belief movement induced by a piece of information. First, it predicts a story-statistic gap: stories are more likely to be successfully retrieved than statistics and are subject to less belief decay. This is because – unlike statistics – stories include qualitative content that is related in meaning to the scenario itself. Second, our model predicts that increasing cue-target similarity by adding elements to a cue that are related in meaning to a given piece of information increases recall likelihood and decreases belief decay when presented with that cue. Third, the model implies that increasing the similarity between a cue and non-target information compromises the likelihood of successful recall and worsens belief decay. Specifically, a non-target story that is related in meaning to the target cue makes it more likely that it will accidentally be retrieved when prompted for a target scenario, thereby increasing interference.

We develop experimental tests of the model's core predictions. First, we consider the core mechanism from our model driving the story-statistic gap in memory: similarity between the target memory and the cue. Intuitively, a story might be particularly likely to be retrieved when its content is semantically related to the cue. On the other hand, for statistics, there is generally no semantic relationship to cues, since numbers are not typically related in meaning to specific concepts. To investigate the relevance of cue-target similarity, we design a mechanism experiment that manipulates the cue-target similarity by varying the similarity between the cue and the story content. We document an important role of cue-target similarity: the more similar the cue is to the story content, the higher the accurate recall and delayed belief movement. Second, we conduct an experiment to test the model's prediction that a higher similarity of interfering information to the target cue decreases accurate recall. Consistent with this conjecture, we show that higher similarity of irrelevant stories to the target cue decreases recall and the persistence of belief impact of the target information. These results imply that in environments where many similar but conflicting stories circulate, stories lose their edge over statistics as a communication device. Taken together, our mechanism experiments highlight the importance of both cue-target similarity and interference due to the similarity of the cue to non-target information for accurate recall. Our findings suggest that to persistently shape beliefs, effective communication should focus on qualitative features that are strongly semantically associated with the cue.

We conclude by examining the relative importance of two potential margins of selective memory: first, people may fail to retrieve any relevant memories for a given scenario; second, people may successfully recall relevant memory traces but only partially recover the original information content. For example, people may remember that the majority of reviews were positive, but not what exact fraction were positive. To connect

the magnitude of belief impact to different recall patterns, we jointly examine our recall and belief data. We document that conditional on correct recall, there remains virtually no story-statistic gap. These analyses provide an empirical foundation for modeling selective recall as primarily arising from retrieval failures rather than from distortions in retrieved content.

Our work relates to a nascent literature on stories and narratives in economics (Shiller, 2017, 2020; Michalopoulos and Xue, 2021; Andre et al., 2022a,b; Kendall and Charles, 2022; Morag and Loewenstein, 2021; Barron and Fries, 2023; Graeber et al., 2023a,b). This literature mostly focuses on the persuasive effects of narratives in moral or political domains (Bénabou et al., 2018; Eliaz and Spiegler, 2020; Bursztyn et al., 2023a,b; Alesina et al., 2023). A related literature in psychology and management studies the power of stories in influencing people (Fryer, 2003; Monarth, 2014; Bruner, 1987; McAdams, 2011). We add to the literature by (i) comparing the effect of stories to that of statistics over time, and (ii) providing systematic, theory-guided evidence on mechanisms with a focus on the role of qualitative information. Our evidence highlights one mechanism by which narratives are effective: they promote recall and thus more easily come to mind at the time of decision-making.

Our work also ties into a growing literature on the role of experiences, attention and memory in economics (Bordalo et al., 2020a, 2021; Gennaioli and Shleifer, 2010; Bordalo et al., 2020b, 2023b; Malmendier and Nagel, 2011, 2016; Link et al., 2023). The model heavily builds on Bordalo et al. (2023a,c), who provide theoretical frameworks in which agents form beliefs by retrieving experiences from memory based on similarity and interference. Enke et al. (2023) empirically study the role of associative memory for belief formation and show that it can give rise to overreaction to news. In contrast to our focus on the decay of belief impact over time, Enke et al. (2023) examine the extent to which the strength of immediate updating in response to new signals is influenced by the history of previous signals. Afrouzi et al. (2023) experimentally study the role of working memory in forecasting experiments. A series of recent papers now also provide evidence on the role of associative recall in field settings, e.g., in finance (Charles, 2022; Jiang et al., 2022; Kwon and Tang, 2023) and the labor market (Conlon and Patel, 2022). Our paper strongly suggests that people do not continuously update their beliefs every time they receive a piece of information, but instead, they partly construct them on-the-fly, consistent with a growing body of evidence on cue-dependent belief formation (Andre et al., 2022a; Bordalo et al., 2021; Enke et al., 2023; Bordalo et al., 2023a). Our paper differs from previous literature in its focus on how different types of information, statistics versus stories, shape beliefs over time.

More broadly, our work builds on extensive psychology literature on memory (Schacter, 2008; Kahana, 2012; Baddeley et al., 2020). Some previous work in psychology

directly relates to the recall of stories, though with a particular focus on the role of scripts (Brewer and Treyens, 1981; Mandler, 1984; Schank and Abelson, 1977; Heath and Heath, 2007), emotions (Kensinger and Schacter, 2008) and mental imagery (Shepard and Cooper, 1986; Standing, 1973; Shepard, 1967). Bower and Clark (1969) document that students' ability to remember a list of words strongly increases when instructed to create a coherent narrative that contains all of the words.[2] These papers differ from ours in a number of ways. First, they focus on studying the recall of word lists, but do not measure beliefs nor track their evolution over time. Second, they do not compare the dynamics of belief formation based on statistics versus stories. Finally, these experiments do not aim to tightly identify underlying cognitive mechanisms, such as the role of cue-target similarity or interference, which are crucial ingredients for models of cue-dependent memory (Bordalo et al., 2023a,c).

This paper proceeds as follows: in Section 2, we outline a simple model of selective memory that formalizes mechanisms driving differences in the recall of story versus statistics. Section 3 presents experiments which demonstrate the existence and robustness of a story-statistic gap in memory. Section 4 describes our evidence on mechanisms. In Section 5, we provide a decomposition of the story-statistic gap and Section 6 discusses the implications of our findings.

2 A Model of Selective Memory

2.1 Setup

We outline a model of memory that adapts Bordalo et al. (2023c,a) to formalize cue-dependent recall and belief formation for stories and statistics. The model setup mirrors our experimental paradigm. Consider a decision-maker (DM) who learns about the reviews a specific product or venue has received. There is a population of N reviews, each of which is either positive or negative. The DM enters with a uniform prior over the number of positive reviews among N. There are two periods. In the first period, the DM may receive additional information about the reviews of the product, either in the form of a story or a statistic. We define a statistic as a randomly drawn subset n of N that includes k positive and $n-k$ negative reviews. For our baseline setup, we define a story as a statistic of $n = 1$ complemented with additional non-quantitative content, akin to an anecdote about a single review.[3] In the second period, the DM receives no additional

[2]This relates to techniques for memory enhancement, which use visualizations of familiar spatial environments to improve the recall of information, commonly referred to as memory palaces or method of loci (Foer, 2012).

[3]However, note that the addition of qualitative content is, in principle, independent of the sample size of the corresponding statistic, which we explore experimentally in Section 3.5.

information. In both periods, the DM states a belief that a randomly drawn review from N is positive.

Over the course of the experiment, the DM faces three scenarios, each one about a different product or venue. Across these three scenarios, the DM receives one story, one statistic, and once no additional information.

2.2 Similarity and Recall

In canonical models of memory, personal experiences are stored in episodic memory (Kahana, 2012). We assume that the DM's episodic memories are organized in a memory database M. Each element of M is a memory trace m that encodes one experience. A trace is a vector of $F \geq 1$ features with values in $V_1 \times \cdots \times V_F$. Some sets of possible values V_f contain the null value $\varnothing$ indicating the absence of a feature. Recall is cue-dependent, which means that recall is initiated by an external trigger that may be a situation, question or specific event. We represent a cue in the same way as memory traces, as a vector of length F with entries in $V_1 \times V_2 \times \cdots \times V_F$. Given any database of traces and any cue, recall is stochastic and governed by the similarity relationships between the cue and the memory traces. We define a similarity measure over any two traces or cue vectors, x_1 and x_2,

$$S(x_1, x_2) : \prod_{i=1}^{F} V_i \times \prod_{i=1}^{F} V_i \to [0, 1],$$

and require that it is symmetric, increasing in the number of features that share the same value, equals 1 if and only if $x_1 = x_2$, and equals 0 if and only if no feature is shared. The probability of recalling a specific target trace m^* when cued with c is given by

$$r(m^*, c) \coloneqq \frac{S(m^*, c)}{\sum_{m \in M} S(m, c)}. \tag{1}$$

The probability of recall is jointly governed by the *cue-target similarity* in the numerator, i.e. the similarity between the cue and the target trace m^*, and *interference* in the denominator, i.e. the similarity between the cue and all other memories. The likelihood of successfully recalling the target trace increases in the cue-target similarity. On the other hand, higher interference yields a higher likelihood of accidentally retrieving an irrelevant memory trace.

2.3 Memory Traces and Cues

Memory traces. In the baseline experiment, there are three scenarios: a bicycle, a restaurant and a video game, creating three corresponding memory traces. We assume the following vector structure for memory traces: the first dimension identifies the scenario, i.e. $V_1 = \{\text{bicycle}, \text{restaurant}, \text{video game}\}$ in the baseline experiment. The second entry encodes the number of reviews n that the DM learns about, i.e. $V_2 = \{0, \ldots, N\}$, with 0 implying that no additional information was received. The third entry carries the number k of positive reviews among the n reviews provided, i.e. $V_3 = \{0, \ldots, N\}$, with $k \leq n$. All additional entries of a trace represent the non-quantitative content provided through a story. These dimensions encode everything that was mentioned in the qualitative content of the scenario. We refer to this set of features by V^{qual} and do not constrain its structure further. V^{qual} is large enough to encode any possible story across all scenarios, and each feature takes the null value or 1, indicating the absence or existence of that feature.[4] For any given trace m, we refer to a realization of the qualitative features as $V^{\text{qual}}(m)$. Note that, here, we focus on how memory traces encode the content that was explicitly provided in a scenario. In addition, standard memory models assume that episodic memories also encode the context in which an experience was made, such as the time of day or weather. Such context could be readily accommodated by $V^{\text{qual}}(m)$, but because most context is plausibly shared across the experiences associated with different scenarios in the experiment, it is largely irrelevant for our purposes besides introducing a base-level similarity between all traces. This is why our exposition of the trace structure focuses on content rather than context.[5]

Memory traces associated with stories and statistics both contain entries in the first three dimensions. The key difference between them is that stories provide additional non-quantitative content, captured by $V^{\text{qual}}(m)$.

Assumption 1. *A memory trace for a scenario has at least one feature present in V^{qual} if a story was received, but none if a statistic or no additional information was received in the baseline experiment.*

In what follows, we denote the treatment type of a memory trace (statistic, story, or dummy) in superscript and the product type in subscript. Given the above assumptions, a statistic conveying that 3 out of 7 reviews for the scenario bicycle were negative forms the following trace in the memory database:

$$m^{\text{statistic}}_{\text{bicycle}} = \left(\text{bicycle},\ 7,\ 3, V^{\text{qual}}(m^{\text{statistic}}_{\text{bicycle}})\right) \quad \text{with} \quad V^{\text{qual}}(m^{\text{statistic}}_{\text{bicycle}}) = (\varnothing, \varnothing, \ldots)$$

[4] For modeling reasons, all memory vectors share the same dimensionality, i.e. all have $V^{\text{qual}}+3$ entries.

[5] It is possible that encoding depends on the informativeness of the information. We abstract from such differences for the sake of simplicity.

A story about a negative review at a restaurant would enter the database as

$$m^{\text{story}}_{\text{restaurant}} = (\text{restaurant},\ 1,\ 0, V^{\text{qual}}(m^{\text{story}}_{\text{restaurant}}) \quad \text{with} \quad V^{\text{qual}}(m^{\text{story}}_{\text{restaurant}}) \neq (\varnothing, \varnothing, \ldots),$$

i.e. $V^{\text{qual}}(m^{\text{story}}_{\text{restaurant}})$ does not only contain null entries. In particular, $V^{\text{qual}}(m^{\text{story}}_{\text{restaurant}})$ contains some entries that represent non-quantitative attributes of the story; for example, that the food was stale or the waiter was unfriendly. The trace produced in a scenario about a video game where no additional information was provided would be encoded as

$$m^{\text{dummy}}_{\text{video game}} = (\text{video game},\ 0,\ 0, V^{\text{qual}}(m^{\text{dummy}}_{\text{video game}})) \quad \text{with} \quad V^{\text{qual}}(m^{\text{dummy}}_{\text{video game}}) = (\varnothing, \varnothing, \ldots).$$

Cues. Retrieval is triggered by a cue, which in our setup emerges from the question about a randomly drawn review of the product. We formalize the cue as invoking a vector c that contains the scenario identifier as well as a (potentially large) set of *semantic associations* with the scenario, $A(c)$. These are connections and relationships that words or concepts have in common with the scenario based on their meanings (McRae and Jones, 2013). The underlying intuition is that the scenario triggers connected concepts that automatically come to mind. For instance, when reading the word "restaurant," natural semantic associations may be "food," "service," or "atmosphere." We abstain from modeling the structure of semantic associations, which is outside the scope of our paper. Instead, we only make the following assumption.

Assumption 2. *For any scenario s, the prompt to form a belief triggers a vector c_s that includes non-null features in V^{qual} that are denoted with $A(c)$.*

The first two assumptions discipline key concepts like stories and semantic associations without explicitly modeling and constraining their structure. The following assumption is central for our results and conceptualizes a highly intuitive aspect of stories.

Assumption 3. *The non-quantitative content of a story delivered in scenario s contains elements that are semantically associated with the scenario. Formally, there is at least one shared feature between $V^{qual}(m^{story}_s)$ and $A(c_s)$.*

The intuition behind this assumption is that a story is actually *about* the scenario at hand; in a natural, relevant story, at least parts of its content are semantically associated with the scenario. The story relates in meaning to the underlying situation. As a result, the memory trace formed by reading a story includes some features that overlap with what the DM semantically associates with the scenario alone. To illustrate, consider a story in which stale sushi was served at a restaurant. This could be represented in the memory trace by a feature encoding (bad) food. Being prompted with the word

"restaurant" automatically triggers thoughts of food, so that the cue and the story trace share the feature "food."

2.4 Belief Formation

The DM forms a belief about whether a randomly drawn review in a given scenario is positive. We assume Bayesian updating given the information that is presented or recalled, so that distortions are limited to the recall process rather than updating biases. Entering with a uniform prior, the DM (potentially) receives additional information on a scenario in the first period. They form a Bayesian posterior and, at the same time, store a single memory trace m^* that follows the structure outlined above. In the second period, the DM is again asked to state their best belief. Rather than recalling their first-period posterior directly, we propose that the DM remembers their original uniform prior, but may or may not remember the additional information received in the meantime.[6] The prompt generates a cue c, which gives the agent a chance $r(m^*, c)$ to recall the target trace m^* and with it the relevant additional information from the scenario. If any other than the target trace is retrieved, the agent notices their mistake and discards it. Successful recall leads to a Bayesian update in the second period that is identical to the first-period belief, whereas failed recall means the agent reverts to their uniform prior.

Notation. For a given scenario we refer to the true probability of a positive review as $\pi := K/N$. The DM's stated belief in period t is $\hat{\pi}_t$.

Prior beliefs. A uniform distribution over $[\![0, N]\!]$ corresponds to a beta-binomial distribution with parameters N and $\alpha = \beta = 1$, inducing the following prior:

$$K \sim \mathcal{U}[\![0, N]\!] = \text{BetaBinomial}(N, 1, 1) \tag{2}$$

Stated belief absent additional information. Absent additional information or when failing to recall it in the second period, the DM relies on their prior. The experiment implements a scoring rule under which reporting the mean maximizes payoffs:

$$\hat{\pi}^{no\ info} = \mathbb{E}_{prior}[\pi] = \frac{1}{2}$$

Stated beliefs with additional information. The DM forms a Bayesian update from the information that there are k positive among n reviews, drawn without replacement

[6] This can be thought of as the DM coming in with a flat ignorance prior, accumulating information over time, and then trying to remember all past information in the moment they face a decision.

from the population of N total reviews. This signal structure follows a hypergeometric conditional distribution:

$$k|K \sim \text{HyperGeometric}(N, K, n) \tag{3}$$

As beta-binomial and hypergeometric distributions are conjugate priors, beliefs about the remaining reviews follow a beta-binomial distribution with parameters $N-n$, $\alpha' := \alpha + k = 1 + k$ and $\beta' := \beta + n - k = 1 + n - k$:

$$K - k|k \sim \text{BetaBinomial}(N - n, 1 + k, 1 + n - k) \tag{4}$$

The average of this distribution is $(N-n)\frac{\alpha'}{\alpha'+\beta'} = (N-n)\frac{k+1}{n+2}$. The DM maximizes payoffs by reporting the mean of the belief distribution:

$$\hat{\pi}^{info} = \mathbb{E}_{posterior}[\pi] = \frac{k}{N} + \frac{N-n}{N}\frac{k+1}{n+2}$$

The first term reflects the certain component, i.e. the knowledge acquired about the subset of observed reviews, whereas the second term captures the uncertain component, i.e. the expected number of positives among the unobserved reviews.[7]

Recall and belief decay. We formalize belief decay as the absolute value of the difference in beliefs formed in the first and the second period. Note that if recall is successful, beliefs are stable so that $\hat{\pi}_2 = \hat{\pi}^{info} = \hat{\pi}_1$, and if recall fails, then $\hat{\pi}_2 = \hat{\pi}^{no\ info}$. The expected second-period belief conditional on period 1 is hence $\mathbb{E}[\hat{\pi}_2 \mid \hat{\pi}_1] = r(m^*, c)\hat{\pi}_1 + (1 - r(m,^* c))\frac{1}{2}$. Belief decay is governed by the probability of recall scaled by the distance of the first-period belief to the prior:

$$\mathbb{E}[|\hat{\pi}_2 - \hat{\pi}_1| \mid \hat{\pi}_1] = (1 - r(m^*, c)) \cdot |\frac{1}{2} - \hat{\pi}_1| \tag{5}$$

2.5 Predictions

Our simple framework makes three main predictions that guide our empirical analysis. The first one establishes the existence of a story-statistic gap. All derivations and additional predictions are relegated to Appendix H.

Prediction 1. *(Story-Statistic Gap.) The likelihood of successful recall is higher for stories than for statistics. Conditional on first-period beliefs, belief decay for stories is lower than for statistics.*

[7] Note that the expected share of positive reviews among unobserved reviews, $\frac{k+1}{n+2}$, is what we would have obtained by a simple application of the rule of succession.

Intuitively, recall of stories is more likely than that of statistics because the additional, non-quantitative content is semantically associated with the scenario (per Assumption 3), and thus stories exhibit a higher cue-target similarity than statistics. A higher likelihood of successful recall induces less belief decay in expectation.

The next two predictions concern the building blocks of the recall mechanism in Equation (1): cue-target similarity and interference. As to the former, we obtain the following implication on the impact of changing the scenario in a way that increases the number of shared features.

Prediction 2. *(Cue-Target Similarity.) Changing the cue to invoke semantic associations that have a larger overlap with the target memory trace raises cue-target similarity. This increases the likelihood of successful recall and decreases belief decay.*

Next we consider the role of interference. Interference is governed by the similarity of non-target memories m to the target cue c, $S(m, c)$. The higher the similarity between the qualitative content of non-target traces and a given target scenario, the more pronounced are forgetting of the target trace and belief decay. To illustrate, assume that there are two scenarios, one about a food truck and one about an amusement park. Consider a story provided in the amusement park scenario that is either about the rides or about the food consumed in the park. The latter story is naturally more closely related in meaning to the other scenario (food truck) than the former story. The similarity between the given cue (food truck) and non-target traces (food in the amusement park) induces a higher probability of accidentally retrieving the memory created for the amusement park in recall about the food truck. In terms of the model, this is reflected in an overlap between $V^{\text{qual}}(m_p^{\text{story}})$ and $A(c_q)$.

Prediction 3. *(Interference.) All else equal, increasing the similarity between a story in scenario p and a cue for another scenario q decreases the likelihood of successful recall and increases belief decay in q.*

3 The Story-Statistic Gap in Memory

3.1 Baseline Design

Our baseline study design is guided by the following objectives: (i) panel data on beliefs that allows us to study the evolution of beliefs over time without new information arriving in the meantime; (ii) a measure of immediate updating that captures any differences in the effects of stories and statistics that are not memory-related; (iii) a naturalistic setting in which information both in the form of statistics and stories is common; and (iv)

an incentive-compatible belief elicitation. Table A.8 provides an overview of all experiments.

Task structure. There are three different hypothetical scenarios, each one about some product or venue.[8] Any given product or scenario has received an overall number of reviews, with each review being either positive or negative. For every scenario, participants' task is to guess whether a randomly selected review is positive. To fix prior beliefs, we truthfully inform them that the actual number of positive reviews would be randomly drawn from a uniform distribution, independently for each scenario, inducing a flat prior. For each scenario, participants then receive either a piece of additional information or no additional information, and are subsequently asked to state their guess.[9]

Timing. In our experiment, we elicit beliefs twice: once immediately upon receiving the information (condition *Immediate*) and once one day later (condition *Delay*). Our main outcome of interest are respondents' incentivized beliefs about the likelihood that a randomly selected review is positive.[10] Below we provide an example of a belief prompt:

> *Out of all the 19 reviews, another review was randomly chosen, where each of the 19 reviews was equally likely to be selected. What do you think is the likelihood (in %) that this review is positive?*

Stories versus statistics. We vary the type of additional information participants are exposed to within-subject and across scenarios. For each scenario, participants receive either statistical information (condition *Statistic*), anecdotal information (condition *Story*) or no further information. Randomization is blocked such that across scenarios, each individual receives one story, one statistic and once no additional information. Moreover, the order of scenarios is randomized and each individual receives one positive signal and one negative signal.[11]

[8] We chose hypothetical scenarios to prevent that relevant additional information can be gathered outside of the experiment.

[9] We included the no information treatment as it adds a natural additional source of uncertainty, namely uncertainty about whether the respondent actually received relevant information in a scenario. Moreover, the treatment allows us to verify that our respondents have understood the treatments by checking whether, absent any additional information, they state a belief of 50%.

[10] The belief elicitation is incentivized using a binarized scoring rule (Hossain and Okui, 2013) with a prize of \$30. The precise payment formula is as follows: Probability of winning \$30 (in percent) $= 100 - 1/100$ (estimate (in percent) - Truth)^2, where truth = 100 if the randomly selected review is positive, and 0 if not. The binarized scoring rule has been shown to be incentive-compatible, even in the presence of risk aversion. Danz et al. (2022) document that empirically, the binarized scoring rule can lead to systematic bias in reported beliefs. Notice that, even if such bias were present in our experiment, it would not compromise our identification which relies on the comparison of beliefs between *Immediate* and *Delay* for stories and statistics. Moreover, all of our findings are supported by evidence on recall, which is immune to the concern about scoring rules.

[11] Appendix E provides details on the implementation of the randomization.

We conceptualize statistics as quantitative information about many reviews. In contrast, we define stories as quantitative information about a single review coupled with qualitative content. Thus, stories and statistics differ along two margins in our baseline setup: first, statistics describe multiple data points, while stories are about only one data point, e.g., an individual experience. The second difference is the presence of qualitative content.[12]

Our design closely adheres to this basic taxonomy. Statistical information is communicated as the number of positive reviews for a randomly selected subsample of the population. The fraction of positive reviews is randomly determined, creating variation in the extremity of statistics. Below is an example of how statistical information is communicated:

> *13 of the reviews were randomly selected. 4 of the 13 selected reviews are positive, the others are negative.*

A story provides information about whether a single randomly selected review is positive or negative, plus a qualitative description of that review. The description consists of six to seven sentences recounting the experience underlying the review. We randomize the direction of the statements made in the text between subjects. For our main analysis, we focus on stories in which the direction of the text statements matches the overall review rating. Below is a shortened example of a story accompanying a negative review about a restaurant:[13]

> *One of the reviews was randomly selected. The selected review is negative. It was provided by Justin... The raw fish looked stale and the sushi rolls were falling apart on the plate... The service was poor: his waiter was rude, not attentive and the food was served after a long wait... As they left the restaurant, Justin was very annoyed and thought to himself "I definitely won't be back!"*

A notable feature of stories is that they cannot be easily accommodated in a Bayesian belief updating framework because the informational content of qualitative statements cannot be quantified in a fully objective way.[14] For instance, in the example above, the qualitative description of the food arguably allows participants to infer that other reviewers may have had similar experiences. Because we cannot determine the normatively optimal Bayesian inference from such qualitative information, we rely on our *Immediate*

[12]Note that in principle, qualitative content could also be added to statistical information. While we maintain that a natural distinction between stories and statistics is that they tend to differ in sample size, Section 3.5 explores the role of adding qualitative content to statistics of fixed sample sizes larger than one.

[13]Appendix D.1 reproduces all stories from the baseline experiment.

[14]In Section 3.4.1 we provide evidence from a setting in which we can cleanly compare belief movements to a Bayesian benchmark.

belief measurement to capture how informative participants *perceive* each story to be – including its qualitative statements.[15] Note that this approach is also not reliant on the assumption that people form their beliefs in accordance with Bayes' rule, which may be commonly violated in practice (Enke and Zimmermann, 2019; Graeber, 2023; Enke, 2020; Martínez-Marquina et al., 2019; Hartzmark et al., 2021; Ba et al., 2023).

Recall elicitation. To provide direct evidence on recall of the additional information about product reviews received in the baseline survey, we asked our respondents the following unincentivized open-ended survey question:[16]

> *Please tell us anything you remember about this product scenario. Include as much detail as you can. Most importantly, please describe things in the order they come to mind, i.e., the first thought first, then the next one etc.*

Hand-coding scheme. To analyze the unstructured text data, we design and implement a hand-coding scheme (see all details in Appendix F). The hand-coding scheme records whether respondents mention the direction and type of information they encountered, and whether they correctly remember these characteristics. It also captures additional features, such as whether (i) respondents in the *Story* condition mention qualitative features, (ii) whether respondents correctly recall the exact statistical information, and (iii) whether respondents recall the belief they stated in the baseline survey. To ensure high quality of the hand-coded data, we proceed as follows. First, we instruct three research assistants on the coding scheme and conduct a series of practice rounds with them. Second, each open text response is independently coded by two of the research assistants. Any potential conflicts are resolved by the third research assistant. We find that the inter-rater reliability is high: for correct recall of type and direction, we find agreement in 94% of the cases.

Incentives. Participants were informed in advance that the survey consisted of two parts, with one day in between. We also told participants that the information they receive would be relevant for payoffs one day later. Participants were truthfully informed that the computer would randomly select 10% of participants to receive a bonus payment that would be based on their responses.[17] To avoid hedging between similar questions in the two parts, one of the three scenarios and one of the two parts for that scenario

[15] Our approach can therefore also account for possible differences in the credibility of the information provided in the *Statistic* versus the *Story* treatments.

[16] We randomized the order of the belief and recall elicitation in the follow-up survey. In additional studies that replicate our baseline findings, we include structured incentivized recall tasks instead of the open-ended question and show that they yield very similar results (see Section 3.4.1).

[17] We paid out close to $15,000 in bonuses across all of our 11 data collections.

(immediate belief, delayed belief) were randomly selected to count for the bonus payment.

Comprehension checks. We implemented an attention check as well as extensive control questions to verify participants' understanding of the instructions. Participation in the survey required passing an attention check and answering all control questions correctly within the first two trials. These control questions ensure high levels of understanding of the payoff incentives as well as the signals and prior distribution of draws.

3.2 Data

Sample. We collected data for the baseline experiment on September 8 (baseline) and September 9 (follow-up) 2022. We recruited participants via Prolific, a survey provider commonly used in social science research (Peer et al., 2022). The average duration of the survey was about 9 minutes for the baseline, and 5 minutes for the follow-up. For the baseline, participants received a completion payment of $1.55 and for the follow-up they received 90 cents.

1,500 respondents completed wave 1 of our experiment. Out of those, 1,437 met the inclusion criteria and were invited for the follow-up survey. 1,035 then completed the follow-up survey. After the pre-specified sample restrictions,[18] our final sample consists of 985 participants, corresponding to a completion rate of 69 percent. Given that the key treatment variation is within-person, the attrition rate is not a threat to the internal validity of our findings. For completeness, we report analyses on attrition rates in Appendix Table A.13.

Pre-registration. All experiments in this paper were conducted online and pre-registered on AsPredicted. The pre-registrations include the experimental design, hypotheses, analysis, sample sizes, and exclusion criteria. A link to each pre-registration is provided in Table A.8. The full set of instructions can be found on the following link: `https://raw.githubusercontent.com/cproth/papers/master/SSM_instructions.pdf`.

3.3 Baseline Results

Beliefs. As pre-registered, we start by analyzing stories with content that is consistent with the overall review rating being positive or negative. The top panel of Figure 1 and Table 1 show the average belief impact in *Immediate* and *Delay*, pooling the data across scenarios and individuals. Belief impact is the signed distance between a stated belief

[18] We pre-specified the exclusion of respondents who indicated having written down the information they received and those updating in the wrong direction in response to statistics.

and the prior (50%). For ease of exposition, we reverse-code the belief impact whenever the additional information implied a downward update, i.e., belief impact is signed in the direction of the rational update. Beliefs in *Immediate* serve as a benchmark that captures any difference in the effect of stories and statistics that is not related to memory.

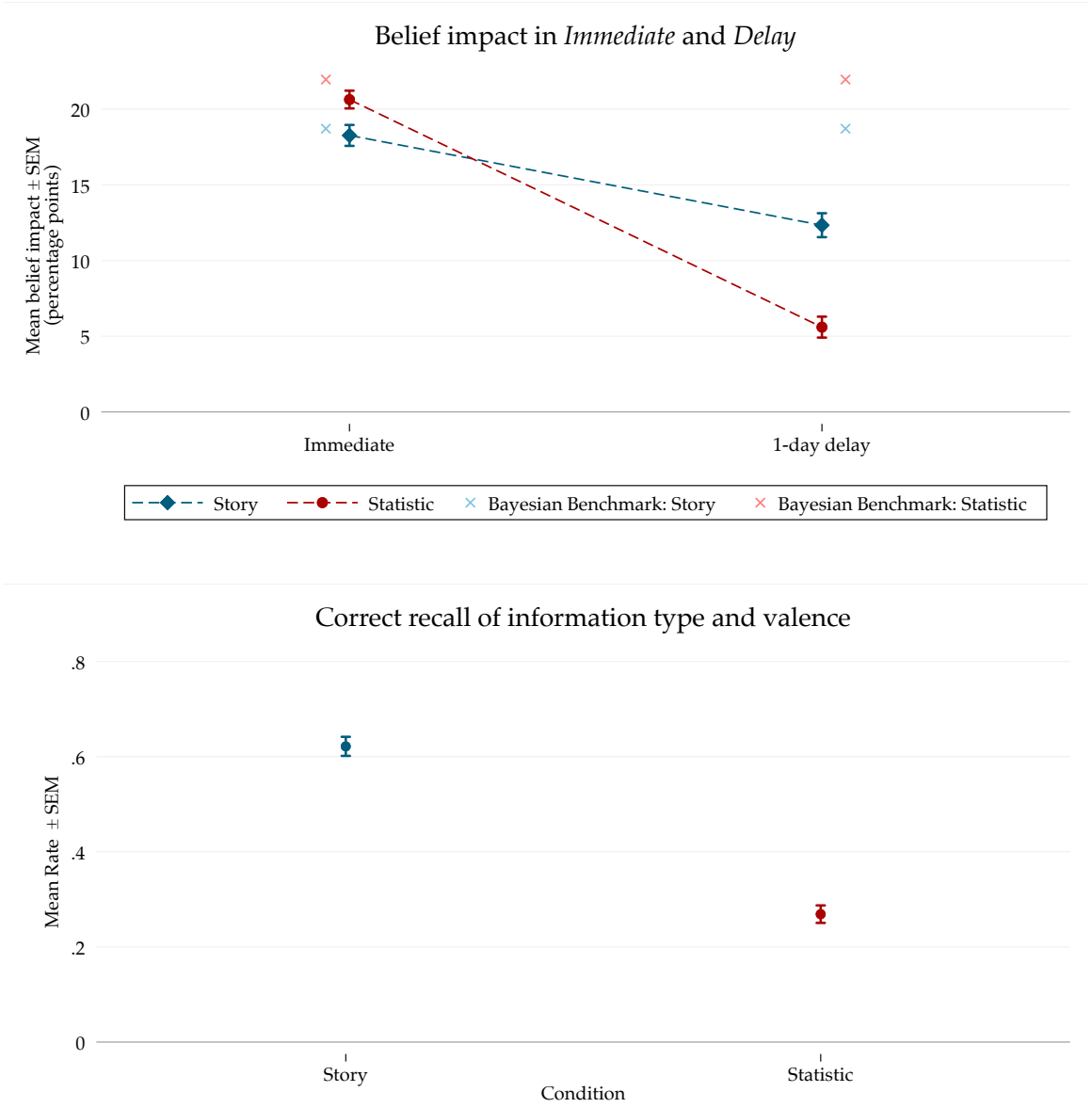


Figure 1: The story-statistic gap in the baseline experiment (984 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The red markers refer to statistics, while the blue markers refer to stories. The light red markers display the average Bayesian benchmark for statistics (21.95 p.p.), while the light blue markers illustrate the average Bayesian benchmark for stories (18.71 p.p.). The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. Whiskers indicate one standard error of the mean.

The top panel of Figure 1 reveals that, in line with our hypothesis, the decay in belief

impact over time is substantially lower for stories than statistics. This is confirmed by column (3) of Table 1. The difference-in-differences estimate of belief impact between *Immediate* and *Delay* is highly significant ($p < 0.01$).

We next consider point estimates of the belief impact in *Immediate*. Average belief impact in *Immediate* is larger for *Statistic* than for *Story*. On average, beliefs moved by 20.63 p.p. (s.e. 0.59) for *Statistic* and by 18.26 p.p. (s.e. 0.69) for *Story*.[19] By contrast, for the *Delay* condition, the top panel of Figure 1 reveals that mean belief impact after one day is substantially more pronounced for *Story* than for *Statistic*. On average, belief impact was 5.60 p.p. (s.e. 0.69) in *Statistic* and 12.33 p.p. (s.e. 0.79) in *Story*. This divergence in belief impact in *Delay* is significantly different from zero ($p < 0.01$). Appendix Figure A.11 underscores these patterns in the cumulative distribution functions of belief impact in *Immediate* and *Delay*, separately for stories and statistics.[20]

Recall. It is conceivable that it may take some time for information to "sink in", and that the beliefs in *Immediate* are elicited before the information has been fully processed.[21] In that case, using the immediate belief as a benchmark may not adequately capture the maximal belief update. We address these concerns using direct data on accurate recall of the provided information.

To study recall, we examine the fraction of respondents who correctly recall both the type and the direction of the information they were provided. The bottom panel of Figure 1 shows that correct recall is significantly higher for stories than for statistics ($p < 0.01$). Average correct recall is 62.15 percent for stories and 26.90 percent for statistics. This suggests that information delivered in the form of stories is more easily retrieved than statistical information. Moreover, open-ended data reveals several other striking features: (i) A large fraction of respondents (44.91%) mention qualitative features from the story without specifically being prompted to do so; (ii) a very small fraction of respondents (1.32%) correctly recall and indicate the statistic they received; and (iii) only a negligible fraction (4.23%) mention the posterior belief they stated in the baseline wave.

[19]The immediate belief impact is close to the (average) Bayesian benchmark for both statistics (22.0 p.p.) and stories (18.7 p.p.). Note that for stories, we only consider the quantitative information contained in the review to compute the Bayesian benchmark, i.e. we do not factor in the potential effect of the qualitative content provided. The experiment reported in Section 3.4.1 provides evidence from a setting where the Bayesian benchmark is well-defined also for stories.

[20]The figure also highlights that there is substantial heterogeneity in perceived informativeness of the story treatment as measured in the immediate condition. This likely arises from differences in the way respondents interpret the qualitative information provided in the story.

[21]In addition, it is conceivable that there are differences between stories and statistics in the extent to which participants rehearse the information they were provided with.

Table 1: The story-statistics gap in memory

	Dependent variable:			
	Belief Impact			Combined Recall
Sample:	Immediate (1)	Delay (2)	Pooled (3)	Consistent (4)
Story	-2.37*	6.73***	-2.37**	0.35***
	(1.23)	(1.48)	(1.01)	(0.03)
Delay			-15.0***	
			(0.90)	
Story × Delay			9.10***	
			(1.28)	
Control Mean	20.63	5.60	20.63	0.27
Observations	1168	1168	2336	1168
R^2	0.54	0.52	0.43	0.65

Notes. This Table uses responses from the *Story* and *Statistic* condition. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Story* takes value 1 for respondents who received a story for a given product, and zero otherwise. *Statistic* takes value 1 for respondents who received a statistic for a given product, and zero otherwise. Columns (1), (2) and (4) include respondents who received consistent stories. Column (3) pools *Immediate* and *Delay*. Columns (1) to (3) display results on belief impact. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Column (4) displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Heterogeneity by extremity of immediate update. Figure 2 illustrates the heterogeneity of delayed belief impact and correct recall by the extremity of the immediate update. The figure showcases that there is little heterogeneity in correct recall by the extremity of the immediate belief update. For all levels of immediate updating, delayed belief impact and correct recall are higher for stories than for statistics.

Result 1. *There is a story-statistic gap in memory: following a delay of one day, stories have a stronger effect on beliefs than statistics, even though statistics have stronger immediate effects, on average. Recall accuracy is substantially higher for stories than for statistics and does not depend on the strength of the immediate update.*

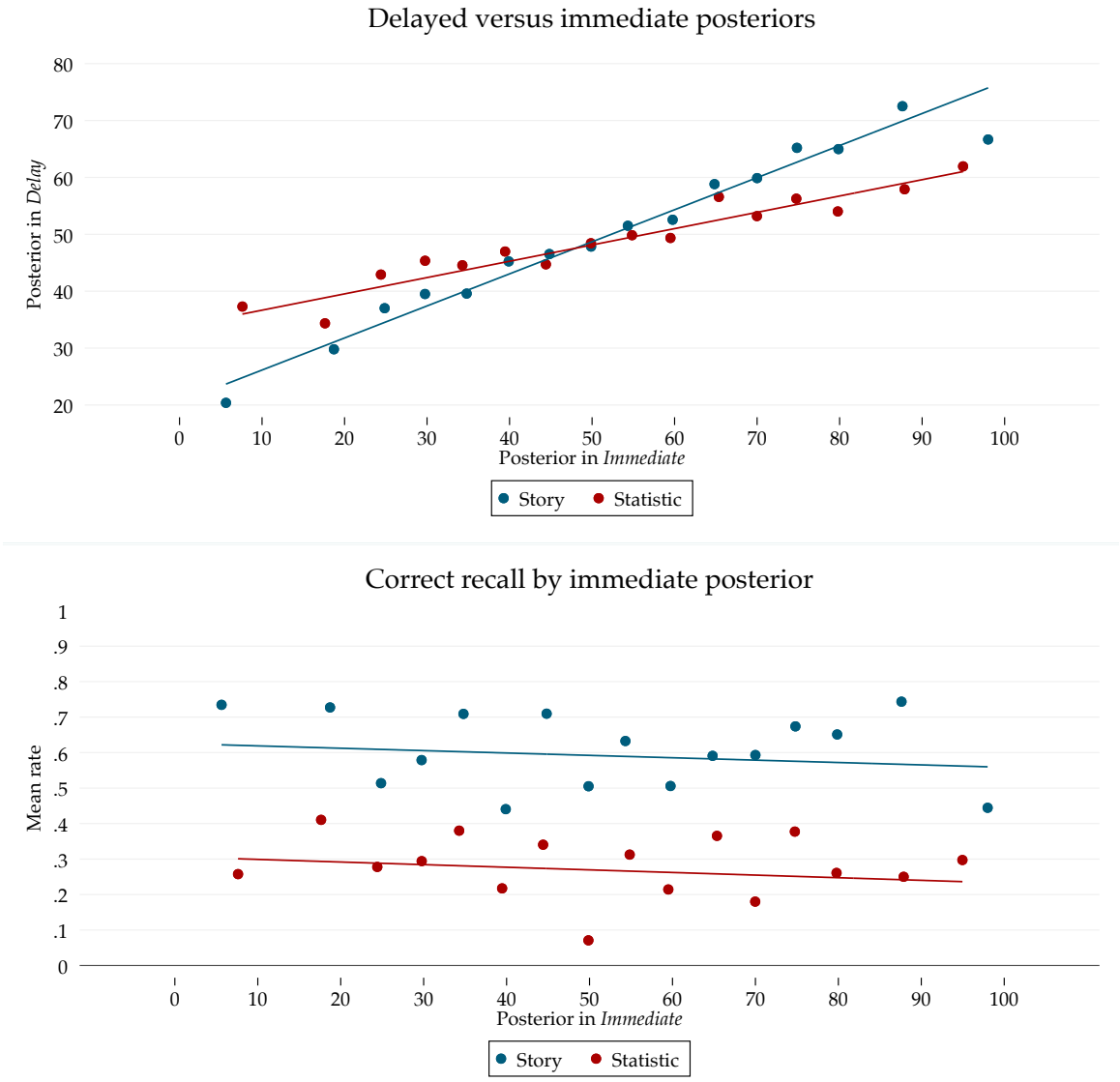


Figure 2: Heterogeneity by extremity of immediate update in the baseline experiment (984 respondents). The top panel displays binned scatterplots regressing beliefs in *Delay* (y-axis) on beliefs in *Immediate*, separately for conditions *Story* and *Statistic*. The bottom panel displays binned scatterplots regressing correct recall of the type and direction of information they received in the baseline survey in *Delay* (y-axis) on beliefs in *Immediate*, separately for conditions *Story* and *Statistic*. The red dots and line illustrate beliefs and recall for statistics, while the blue dots and line illustrate beliefs and recall for stories.

3.4 Robustness

3.4.1 Experiment with Uninformative Qualitative Content

Design. A possible concern with our baseline design is that, as discussed above, we cannot compute Bayesian benchmarks for the Story treatment. Respondents may interpret the qualitative content to be informative above and beyond the quantitative information. To deal with this concern, we conduct a robustness experiment which is identical to our baseline experiment except that it explicitly tells respondents that the qualitative con-

tent of the stories is uninformative, hence allowing for the computation of a Bayesian benchmark. We directly confirm that respondents pay attention to and understand this instructions using an additional comprehension question. Appendix A.1 provides additional details.

Sample. We recruited 1,000 respondents for the baseline survey. 912 respondents qualified for the follow-up survey. After the pre-specified sample restrictions, our final sample consists of 714 respondents, corresponding to a completion rate of 78 percent.

Belief movement. Appendix Figure A.1 confirms our baseline finding: The decay in belief impact over time is significantly lower for stories than statistics ($p < 0.01$), as illustrated by column (3) of Appendix Table A.1.

We next consider point estimates of the belief impact in *Immediate*. Average belief impact in *Immediate* is larger for *Statistic* than for *Story*. On average, beliefs moved by 21.91 p.p. (s.e. 0.49) for *Statistic* and by 19.21 p.p. (s.e. 0.51) for *Story*. The immediate belief impact is close to the average Bayesian benchmark for both statistics (22.07 p.p.) and stories (18.68 p.p.).[22]

For the *Delay* condition, belief impact is 8.57 p.p. (s.e. 0.62) in *Statistic* and 10.52 p.p. (s.e. 0.66) in *Story*. This divergence in belief impact in *Delay* is significantly different from zero ($p < 0.01$). These findings underscore that respondents in the *Story* condition hold delayed beliefs that are significantly closer to the Bayesian benchmark.

While this difference is still highly significant it is less quantitatively large compared to our baseline experiment. This likely arises from the fact that we explicitly told respondents that the anecdotal details are not informative beyond the information contained in the quantitative reviews.

3.4.2 Incentivized structured recall

To complement our open-ended measure of recall from the baseline experiment with an incentivized measure, we use a structured recall task. We ask respondents to indicate whether they (i) received information about a single review, including some additional anecdotal details about the reviewer and their experience with the product, (ii) multiple

[22]The difference between the two benchmarks might seem surprisingly small. Appendix G shows that after observing a single story out of 17 total reviews, the Bayesian belief movement is 0.187, identical to our sample average. This is because stories are always positive or negative, and therefore extreme draws. On the other hand, most statistics that respondents learn about are rather moderate. When exactly half the observed reviews are positive, the Bayesian belief movement is 0. Appendix G shows that even with intermediate draws, this movement is modest. Averaging over all draws leads to the relatively modest average belief impact seen in our sample. Put simply, all stories are extreme, while most statistics are moderate.

reviews, (iii) no information or (iv) don't know.[23] Unless respondents indicate that they did not receive any information about this product, we additionally ask them to indicate whether the information they received was positive or negative.[24] Respondents are told that if they correctly recall the information they received, they will obtain an additional bonus of $5. To circumvent hedging motives, either beliefs or recall were randomly selected for payment, and one question was randomly chosen to determine the bonus. The bottom panel of Appendix Figure A.1 shows that correct recall is significantly higher for stories (69 percent) than for statistics (32 percent).

3.4.3 Willingness to pay

To examine whether the information provided in our reviews would likely affect consumer behavior, we elicited a hypothetical willingness to pay measure after the belief elicitation. Table A.2 in Appendix A.1 shows that the *Story* and *Statistic* treatments both significantly increase willingness to pay compared to the no information treatment, both in the *Immediate* and *Delay* conditions. Given that the initial differences in willingness to pay are highly significant, this data does not lend itself to cleanly study memory mechanisms. Nonetheless, the additional measure suggests that the provided information is related to hypothetical willingness to pay.

3.4.4 Robustness to Design Features

In Appendix A we provide a series of additional robustness checks. In Appendix A.2 we show that manipulating the valence of story content only has very little effect on accurate recall. In Appendix A.3 we examine heterogeneity of our findings by positive versus negative information. Appendix A.4 showcases the robustness of the story-statistic gap to different combinations of non-target information. In Appendix A.5 we examine how the size of the story-statistic gap depends on the number of products. Finally, Appendix A.6 confirms the robustness of our findings to using different belief elicitation formats.[25]

[23]Respondents are told that if they choose "don't know", one of the other options will be randomly chosen to determine their payoff.

[24]To keep the elicitation for stories and statistics as comparable as possible, we asked respondents who indicated having received multiple reviews whether the majority of reviews was positive or negative, while respondents indicating having received a single review were asked whether the single review was positive or negative. In this elicitation respondents can again select "don't know".

[25]In a previous version of this paper we conducted an additional experiment that varied the similarity between different scenario names (Restaurant A, Restaurant B, Restaurant C). This evidence shows that increasing cue similarity decreases forgetting significantly. Because this design is not tightly connected to the conceptual framework anymore, this evidence is described in Appendix B.1.

3.5 Statistics with Qualitative Content

In line with our conceptualization of stories versus statistics, the story-statistic treatment variation from our baseline experiment changes both the number of reviews and varies the presence of additional qualitative content. In our model of selective memory, however, it is only the latter dimension, qualitative content, that drives the story-statistic gap. To isolate the role of qualitative features, we conduct an additional experiment, in which information-free qualitative features are added to statistics.

Design. The incentives and basic setting are identical to the experiment described in Section 3.4.1. The only difference relative to this design concerns the information respondents receive: for each product, participants receive either statistical information (condition *Statistic*), statistical information with an uninformative anecdote about one review (condition *Statistic with qualitative content*), or no further information. In the *Statistic with qualitative content* treatment, respondents are told that they receive additional details about one of the reviews. We employ the same reviews as in the baseline experiment and always provide respondents with anecdotal details that are consistent with the direction of the statistic.

Sample. We recruited 1000 respondents for the baseline survey. 906 respondents qualified for the follow-up survey. After the pre-specified sample restrictions, our final sample consists of 673 respondents, corresponding to a completion rate of 74 percent.

Results. The top panel of Figure 3 as well as Table A.10 confirm that the decay in belief impact over time is significantly lower for statistics with qualitative content than statistics without qualitative content ($p < 0.01$).[26] The bottom panel of Figure 3 shows that correct recall is significantly higher for statistics with qualitative content than for statistics alone. Average correct recall is 58.10 percent for statistics with qualitative content and 21.40 percent for statistics ($p < 0.01$). These findings confirm our model predictions and underscore the importance of qualitative content in driving the story-statistic gap in memory.

Endogenous qualitative content. In Appendix A.7 we report the design and results of a related experiment. Instead of exogenously adding qualitative content to statistics, we

[26] Immediate impact for statistics with qualitative content is larger than for statistics without qualitative content, even though all respondents passed an attention screen verifying that they understood that the qualitative information carries no additional information. This possibly arises from the qualitative stimuli enhancing the process of mental simulation (Bordalo et al., 2023a).

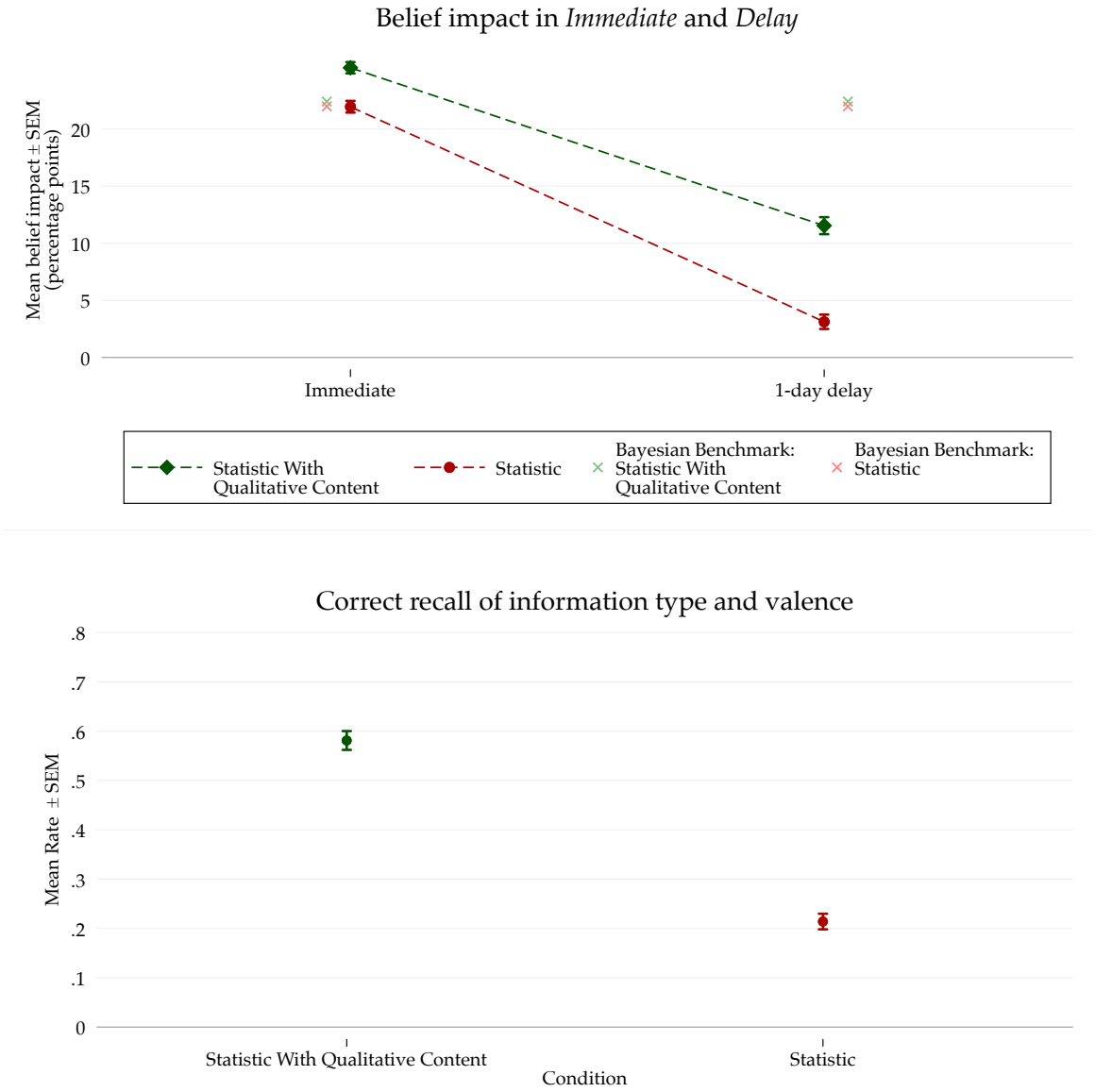


Figure 3: Gap in belief impact and recall for statistics with and without qualitative content (673 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The red markers illustrate belief impact and recall for statistics, while the green markers illustrate belief impact and recall for statistics with qualitative content. The light green markers display the average Bayesian benchmark for statistics with qualitative content (22.41 p.p.), while the light red markers illustrate the average Bayesian benchmark for statistics (21.98 p.p.). Whiskers indicate one standard error of the mean.

asked respondents to imagine and describe a typical review in light of statistical information provided on the belief elicitation screen in *Immediate*. Our results indicate that prompting respondents to come up with a typical review when provided with a statistic increases delayed belief impact and improves recall accuracy, even though immediate

updating remains unaffected by the prompt. Put differently, asking participants to add fictional qualitative features to a statistic on their own slows the decay of its belief impact, further highlighting the crucial role of qualitative content for the story-statistic gap.

3.6 Interpreting the Story-Statistic Gap

There are many differences between stories and statistics, and our baseline evidence leaves open which of these are responsible for the story-statistic gap. In the following, we provide a brief overview and discussion.

Engagement with additional information and processing time. Differences in the processing time of stories and statistics, which may be indicative of the encoding strength, are a plausible mechanism underlying the story-statistic gap. We find that respondents spend somewhat more time processing stories (median of 42 seconds) than statistics (median of 32 seconds). Appendix Table A.9 examines heterogeneity in belief impact and recall by the time spent processing the information. Correlationally, we find small and insignificant heterogeneity in differential belief impact based on initial processing time. Moreover, our mechanism experiments in Section 4 hold the processing time of the target scenario constant, as they only vary similarity relationships between the cue and the informational content of stories.

Emotions and vividness. Research in psychology has established a connection between emotions and memory (e.g., Kensinger and Schacter, 2008). Our evidence on the direction of story content (Appendix A.2) suggests that while stories with more consistent qualitative features are recalled at somewhat higher rates than stories with mixed and neutral qualitative story content, these differences are relatively small, especially compared to the large differences in recall between stories and statistics. Moreover, while emotions plausibly play a role in driving the baseline story-statistic gap, the bulk of our mechanism evidence focuses on the features of cue-dependent memory, which allows us to hold emotions fixed.

Outside memories and sample. Respondents do not enter the experiment with a blank slate but bring in an outside database of memories. This existing database will contain both stories and statistics to some extent, potentially affecting memory of different types of information. Observing information in the experiment (either in the form of a story or a statistic) might trigger the recall of such outside memories. These recollections from outside the experiment might in turn influence beliefs elicited in the

experiment. In other words, outside memories might have a *resonance effect*, akin to information resonance in Malmendier and Veldkamp (2022). Our experimental instructions clarify that the only relevant information for the experimental tasks is the one provided within the experiment. Furthermore, it is important to note that such effects, if present, would already influence beliefs in the immediate condition. Hence, in order for resonance effects to explain the differences in the dynamic pattern of belief impact between stories and statistics, one would need to allow for such resonance effects to differ between immediate and delay. Alternatively, information resonance might shape how well the information is encoded and stored. Finally, our mechanism experiments, which hold the target information constant and operate by changing the similarity relationships, are immune to the information resonance mechanism.

4 Mechanisms

Guided by the predictions spelled out in Section 2, we proceed with an analysis of underlying mechanisms. First, we investigate the role of cue-target similarity in Section 4.1. Second, we examine the importance of interference in Section 4.2.

4.1 Cue-Target Similarity

Our model suggests that one key driving force behind the story-statistic gap is the similarity between the cue and the target information. The qualitative content of stories is often related in meaning to the scenario that serves as the cue. As a consequence, stories more easily come to mind than statistics, which, due to their abstract nature, tend to be unrelated to cues. To examine the role of cue-target similarity, we conduct experiments that manipulate the degree of similarity between stories and cues.

Design. The incentives and basic setting are identical to the experiments that allow for the computation of Bayesian benchmarks for stories and statistics. Unlike in the baseline experiments where information types are randomized across scenarios, we here always provide a story in the restaurant scenario. Our key treatment varies, between subjects, the similarity between the cue, i.e., the name of the scenario and the story for that scenario. Specifically, for the restaurant scenario, each respondent receives either a positive or a negative story. The story describes a reviewer's experience in an Italian restaurant, where the experience makes explicit reference to the Italian cuisine of the restaurant.[27] In the other two scenarios, respondents receive a statistic once and once no additional

[27]Appendix D.2 reproduces the stories.

information. We then vary across participants how similar the cue in the restaurant scenario is to the content of the story. We have three conditions where we systematically vary the similarity. In the *High Similarity* condition, the name of the restaurant is *The Italian restaurant "Napoli"*. In *Low Similarity 1*, the name of the restaurant is *An eatery*, while in *Low Similarity 2* the name of the restaurant is *Mr. Jones*. The high similarity cue mentions Italian cuisine and repeats the name of the restaurant as mentioned in the story content, while the low similarity cues have no direct association with Italian food. *Low Similarity 1* is a generic term for a dining establishment, while *Low Similarity 2* is a generic name that does not even reveal that the venue of interest is a restaurant.

Sample. We recruited 1000 respondents for the baseline survey. 912 respondents qualified for the follow-up survey. After the pre-specified sample restrictions, our final sample consists of 627 respondents, corresponding to a completion rate of 69 percent.

Result. Figure 4 and Appendix Table A.11 presents the results from this experiment. The upper panel shows results on immediate and delayed belief movement. The figure illustrates that the decay in belief impact over time is significantly lower for *High Similarity* than *Low Similarity 1* ($p < 0.01$) and *Low Similarity 2* ($p < 0.01$). The comparison of belief decay is straightforward as beliefs in *Immediate* do not differ significantly.

The lower panel of Figure 4 confirms these patterns in the recall data. While accurate recall in *High Similarity* is at 80 percent, it is at only 70 percent and 38 percent in *Low Similarity 1* and *Low Similarity 2*, respectively. These effect sizes are large in magnitude and highly statistically significant ($p < 0.01$). Taken together, these results underscore the quantitative importance of the cue-target similarity mechanism.

Reassuringly, Appendix Figure A.12 shows that there are no significant differences in the belief movement and recall for statistics across the three treatment conditions. This suggests that, as intended, our manipulation only affected recall of the restaurant scenario. Moreover, this pattern allows us to cleanly cast our findings in terms of the story-statistic gap. Increasing cue-story similarity significantly increases the story-statistic gap in memory.

Result 2. *Increases in cue-target similarity significantly increase delayed belief impact and accurate recall.*

4.2 Similarity of Cue to non-Target Information

While the qualitative content of stories tends to give them an edge over statistics due to its natural similarity to the cue, our model also clarifies that qualitative content of non-target traces can inhibit the recall of target traces due to interference. Specifically, a

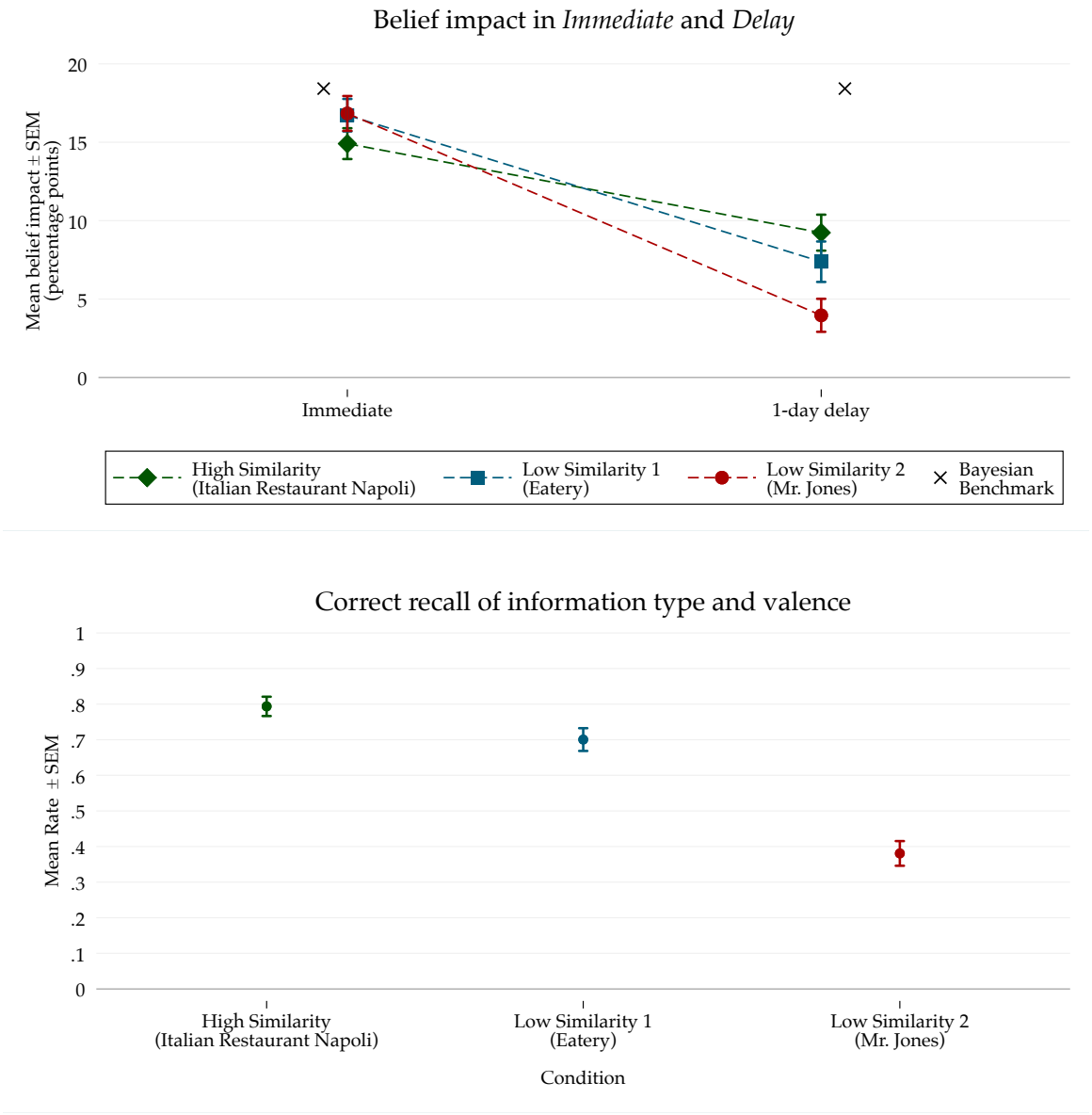


Figure 4: Belief impact and recall in Mechanism Experiment 1 (670 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The green markers illustrate belief impact and recall for *High Similarity*, the blue markers illustrate belief impact and recall for *Low Similarity 1*, while the red markers show belief impact and recall for *Low Similarity 2*. The black markers illustrate the Bayesian benchmark (18.45 p.p.). Whiskers indicate one standard error of the mean.

direct prediction of our model is that a higher similarity between the target cue and non-target stories creates memory interference in the recall of target information and hence reduces belief movement and accurate recall after some delay. To test this prediction, we conduct an experiment with stories only. This experiment directly manipulates the similarity between the cue and a set of non-target stories.

Design. The incentives and basic setting are identical to the mechanism experiment presented in the previous section. We have two treatment conditions that vary, between subjects, the similarity of the target cue to non-target information. All participants learn about three scenarios: a food truck, an amusement park, and a sports stadium. Unlike in our main experiment, respondents receive a story in each of the three scenarios. The target story in both conditions that our analysis focuses on is a positive review about a food truck. The stories about the amusement park and the sport stadium are non-target stories and both feature a negative review. In *Low Similarity*, the three stories are distinct and specific to each cue. The food truck story describes the quality of a hot dog, while the sports stadium and amusement park stories describe features of the stadium and the amusement park, respectively. In the *High Similarity* condition, we keep the target story about the food truck identical to *Low Similarity*, but increase the similarity of the two non-target stories to the cue by modifying the content of the stories. Specifically, in *High Similarity*, the two non-target cues are still amusement park and sports stadium, but now the stories describe hot dogs consumed at these locations. Thus, our treatment fixes the target story and only manipulates the similarity between the two non-target stories and the food truck cue. All other design aspects are identical between the conditions. Appendix D.3 reproduces all stories that we used.

Sample. We recruited 1,000 respondents, of which 670 qualified for the follow-up.[28] After the pre-specified sample restrictions, we have a sample size of 505, corresponding to a completion rate of 75 percent.[29]

Results. The top panel of Figure 5 shows data on the belief impact of the target story in *Immediate* and *Delay*, separately for *High Similarity* and *Low Similarity*. While there is no difference in belief impact in *Immediate*, the slope in belief impact is steeper in *High Similarity* compared to *Low Similarity*, in line with the model prediction. Delayed belief impact is significantly lower in *High Similarity* than in *Low Similarity*.

While average delayed belief impact in *High Similarity* is 5.91 p.p. (s.e. 1.05), it is 9.86 p.p. (s.e. 1.12) in *Low Similarity*. Table A.11 confirms this visual pattern and shows that the difference-in-differences in belief impact (difference in slopes) is statistically significant ($p < 0.01$).[30]

[28]The somewhat larger fraction of respondents not qualifying for the follow-up study can be explained by them failing our pre-specified inclusion criterion of updating in the right direction in *Immediate* in all three scenarios.

[29]The completion rate to the follow-up survey does not differ significantly across treatment groups ($p = 0.62$).

[30]In Appendix B.3 we confirm the robustness of these results on interference using another experiment that featured a different set of cues and stories but overall employed a similar design.

The bottom panel illustrates similar patterns for recall: Among respondents in *Low Similarity*, 48.36 p.p. (s.e. 3.43) correctly recall the information, compared to only 31.16 p.p. (s.e. 2.72) in *High Similarity*. This difference of 17.2 p.p. is statistically significant ($p < 0.01$). This effect size is moderate in size and corresponds to 0.35 of a standard deviation.[31]

Result 3. *Increases in similarity of the target cue to non-target information significantly decreases delayed belief impact and recall accuracy of the target information.*

Implications. This finding has several implications. First, it provides strong evidence for the power of similarity relationships in determining the decay of belief impact and recall accuracy. Second, it delineates the limits of the stickiness of stories in memory. If the memory database contains many stories that are similar to a target cue, retrieval of a target story gets crowded out and it becomes less likely that this story comes to mind. Hence, stories as a communication device lose their edge in environments where similar stories circulate.

5 Decomposing the Story-Statistic Gap

The evidence presented so far leaves some fundamental questions on the processes underlying selective memory unaddressed. One distinction with far-reaching consequences for both theoretical and empirical work is whether distortions induced by memory result from (i) successfully retrieving memories that are subject to partial information loss or (ii) complete retrieval failures. In the following, we provide a decomposition with the purpose of quantifying the importance of these two margins of memory in driving the story-statistic gap.[32]

To examine these ideas, note first that the basic memory retrieval process can have three different possible outcomes: (a) retrieval failure, (b) successful retrieval without information loss, and (c) successful retrieval of a memory trace that is subject to information loss. Each of these retrieval outcomes is associated with a different signature in the decay of belief impact. First, the DM may not retrieve any target memory and beliefs therefore revert to the prior (class *FullDecay*). Such retrieval failure creates a clear benchmark of full belief decay. Second, the DM may correctly recall the full wealth

[31] Forgetting in *Low Similarity* of this mechanism experiment is higher than in our baseline experiment for potentially two reasons: First, the pieces of additional information (three stories) may be more similar to one another than the information provided in the baseline experiment. Second, respondents in this mechanism experiment receive three pieces of information instead of only two pieces of information in the baseline experiment.

[32] The analyses in this section are exploratory in nature and were not pre-registered.

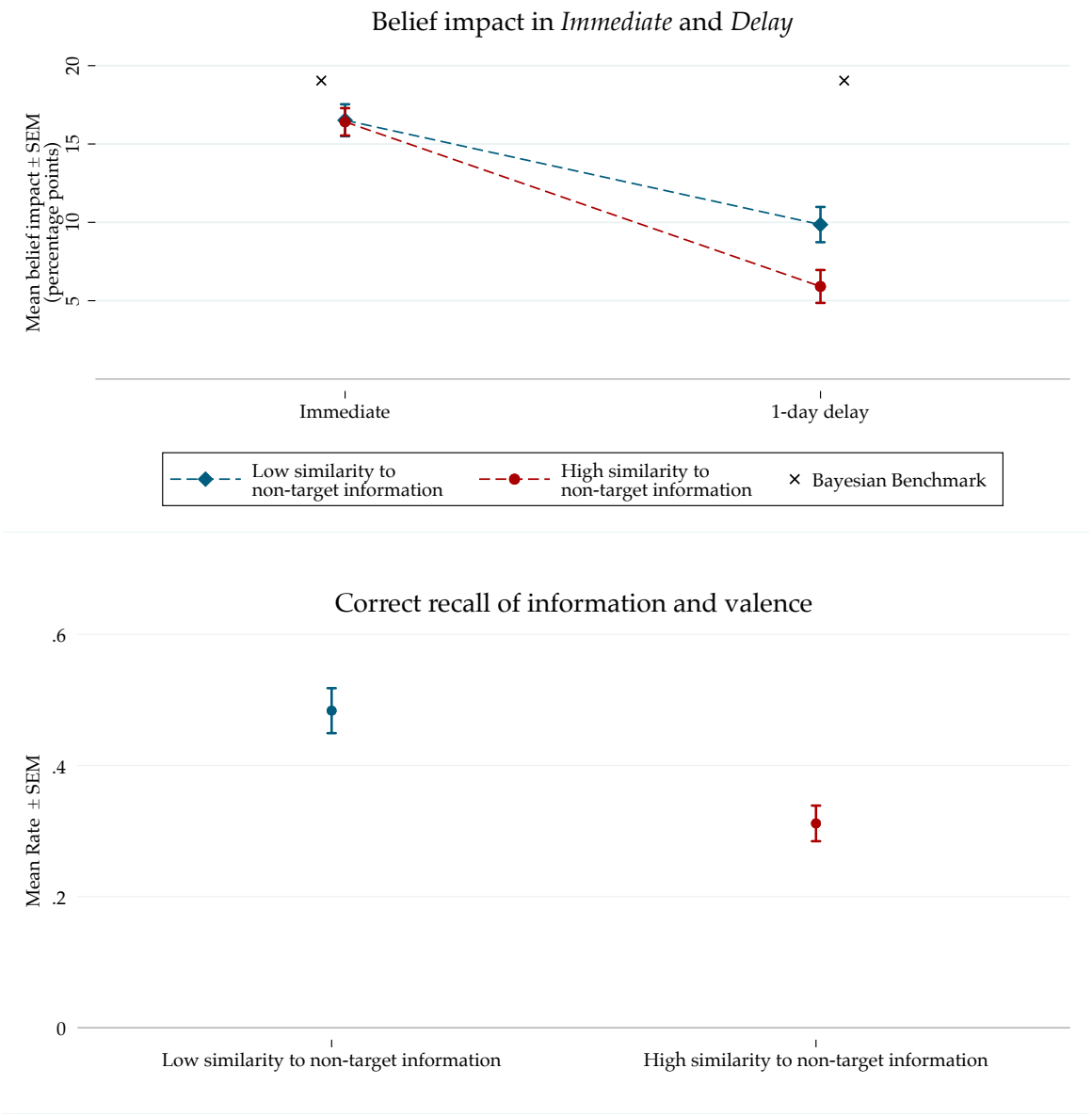


Figure 5: Belief impact and recall in Mechanism Experiment 2 (505 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The blue markers illustrate belief impact and recall for *Low similarity to non-target information*, while the red markers illustrate belief impact and recall for *High similarity to non-target information*. Whiskers indicate one standard error of the mean.

of information contained in a story or statistic (class *ZeroDecay*). In that case, the DM would state their past posterior belief and we would observe no (or very little) decay in beliefs over time. Third, the DM may successfully recall the target memory trace, but the retrieved information is subject to information loss (class *IntermediateDecay*). The corresponding signature in beliefs would be a partial reversion to the prior.

The combination of recall data and the evolution of beliefs allow us to shed some

light on the relative importance of these margins of memory distortions. Specifically, we proceed as follows: We use recall data to identify retrieval failures and test whether this class is, in fact, associated with *FullDecay* in the corresponding stated beliefs. We then turn to the remaining data, which, by construction, only include observations where people correctly remember at least some of the information – information type and direction. We focus on the corresponding belief data to assess the relative shares of *ZeroDecay* and *IntermediateDecay*. The following analysis focuses on our baseline experiment reported in Section 3.

First, recall that the bottom panel of Figure 1 identifies the fraction of beliefs associated with retrieval failure, specifically, a failure to retrieve information (type and direction) across conditions. Following this metric, 38 percent of observations in *Story*, but 74 percent in *Statistic* fail to retrieve relevant information about the target trace. According to the model, these observations should be associated with beliefs that fully revert to the prior of 50%, implying a belief impact of zero (class *FullDecay*). Figure 6 displays the story-statistic gap in belief impact separately for the sample of observations associated with correct and incorrect recall (following the definition of the bottom panel of Figure 1). The average belief impact for observations classified as *FullDecay* indeed reverts to close to zero in *Delay*, as predicted by the model.

Second, turning to observations associated with correct recall of type and direction, we can identify the class of *ZeroDecay* as those that state identical beliefs in *Immediate* and *Delay*. This comprises 37.67 percent (47.66 percent of correct recall observations) of all observations in *Story* and 30.65 percent (56.05 percent of correct recall observations) of all observations in *Statistic*. Note that these figures likely identify a lower bound, because they do not take into account potential measurement noise in beliefs. If people in *ZeroDecay* answer the belief questions with some added random noise, there would be no average belief decay, yet many would state beliefs that differ between the two periods.[33]

Finally, we turn to the remaining class, *IntermediateDecay*, which is associated with correct recall of type and direction, but at the same time features beliefs with *some* intensive-margin information loss by virtue of neither being part of *ZeroDecay* nor *FullDecay*. Above we already classified a substantial lower bound for the class *ZeroDecay*. Figure 6 displays average belief decay among observations associated with correct recall of type and direction. Strikingly, it reveals that there is zero average belief decay in the *Story* condition and a quantitatively minor, only marginally significant decay in the *Statistic* condition. Put differently, conditional on correct recall, we see close to no evidence for belief decay, suggesting a central role of retrieval failures.

[33]We can instead apply a more lenient benchmark than precisely zero decay, but, as will be clear below, this will, if anything, only strengthen our conclusion.

Taken together, this exercise provides a clear conclusion: In our experiments, patterns of selective memory are driven by a failure to retrieve any relevant memory for a given scenario, rather than successful recall with partial information loss.

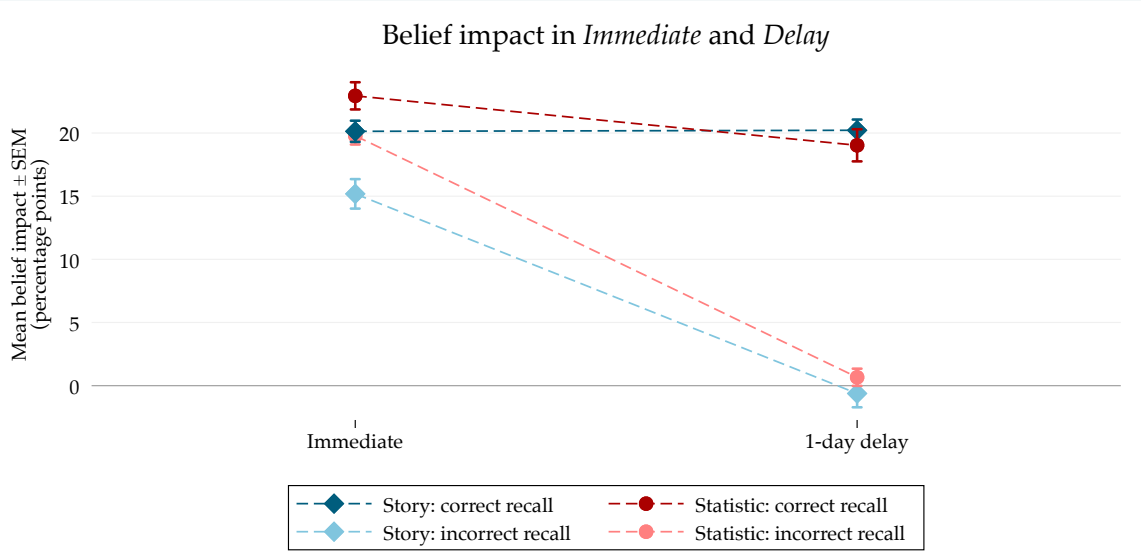


Figure 6: The decay of belief impact by recall accuracy in the baseline experiment (984 respondents). The figure displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The dark blue markers illustrate belief impact for stories with correct recall, while the light blue markers illustrate belief impact for stories with incorrect recall. The dark red markers illustrate belief impact for statistics with correct recall, while the light red markers illustrate belief impact for statistics with incorrect recall. Whiskers indicate one standard error of the mean.

6 Discussion and Conclusion

This paper documents a story-statistic gap in memory. As time passes, the effect of information on beliefs generally decays, but this decay is much less pronounced for stories than for statistics. Using recall data, we show that stories are more accurately retrieved from memory than statistics. We causally show that this pattern is driven by the presence of qualitative content in stories. Guided by a simple model of cue-dependent memory, we experimentally demonstrate the explanatory power of two key forces of memory: cue-target similarity and interference. Our memory decomposition provides striking evidence that retrieval failures appear to be the key driver of the story-statistic gap, rather than partial information loss in retrieved memories.

Stories in the mass media. Our findings have implications for understanding several real-world phenomena. Mass media provides not only facts and statistics, but also relies on anecdotes about individual cases, which provide detailed qualitative information.

Consider allegations about election fraud in the context of the 2020 U.S. presidential election, where some outlets reported stories about individual instances of election fraud, even though these were rare exceptions. Likewise, consider news reporting about welfare fraud where anecdotes about individual cases are abundant in the news media, but stand in stark contrast to official statistics on fraud incidence. For example, Ronald Reagan, beginning with his 1976 campaign, told extreme stories about "welfare queens:"

> *She has 80 names, 30 addresses, 12 Social Security cards and is collecting veterans' benefits on four non-existing deceased husbands. And she's collecting Social Security on her cards. She's got Medicaid, getting food stamps, and she is collecting welfare under each of her names. Her tax-free cash income alone is over $150,000.*

Similarly, consider mass media coverage of immigration. While statistics about low crime rates among immigrants are widely reported by news outlets, extreme stories about immigrants committing severe crimes also regularly hit the headlines.

Our results indicate that stories disseminated in this way can have powerful effects on beliefs as they may come to mind more easily than more representative statistical information. This provides a potential explanation for the emergence of widely documented misperceptions about real-world topics, and for the persistence of these distortions.

Policy communication. Our results also have implications for the communication of statistical information. If policymakers, marketers or leaders aim to convey statistical information effectively, they may wish to complement it with anecdotes to ensure that the information sticks with the audience. For instance, statistical information about economic quantities could be coupled with anecdotal information that is consistent and inherently reminiscent of the embedded statistical information. Moreover, our results suggest that persuaders should factor in the time structure when picking their mode of persuasion: if messaging occurs close in time to the audience's anticipated action, statistics and quantitative facts can be more powerful than stories; yet, as soon as a delay is involved, stories trump statistics.

Avenues for future research. First, it would be desirable to gain a better understanding of the evolution of memory patterns and the story-statistic gap as time delays increase. Second, to shed light on the external validity of our findings, it will be important to assess whether our results are specific to the recall of statistical information, or whether they instead extend to the recall of simple facts devoid of any context. Third, it would be interesting to understand how memory mechanisms affect the virality of different types of information. Finally, future work might examine whether there is also

a gap in the evolution of beliefs between personal and non-personal experiences that is analogous to the story-statistic gap.

References

Afrouzi, Hassan, Spencer Yongwook Kwon, Augustin Landier, Yueran Ma, and David Thesmar, "Overreaction in Expectations: Evidence and Theory," *Quarterly Journal of Economics*, 2023, *138* (3), 1713–1764.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva, "Immigration and Redistribution," *Review of Economic Studies*, 2023, *90* (1), 1–39.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective Models of the Macroeconomy: Evidence from Experts and Representative Samples," *Review of Economic Studies*, 02 2022, *89* (6), 2958–2991.

_ , Ingar Haaland, Christopher Roth, and Johannes Wohlfart, "Narratives about the Macroeconomy," *Working Paper*, 2022.

Ba, Cuimin, Aislinn Bohren, and Alex Imas, "Over- and Underreaction to Information," *Working Paper*, 2023.

Baddeley, Alan, Michael W. Eysenck, and Michael C. Anderson, *Memory*, Routledge, 2020.

Barron, Kai and Tilman Fries, "Narrative persuasion," 2023.

Bénabou, Roland, Armin Falk, and Jean Tirole, "Narratives, Imperatives, and Moral Reasoning," *Working Paper*, 2018.

Bordalo, Pedro, Giovanni Burro, Katie Coffman, Nicola Gennaioli, and Andrei Shleifer, "Imagining the Future: Memory, Simulation and Beliefs," *Working Paper*, 2023.

_ , John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer, "How People Use Statistics," *Working Paper*, 2023.

_ , _ , _ , _ , and _ , "Memory and Probability," *Quarterly Journal of Economics*, 2023, *138* (1), 265–311.

_ , Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer, "Memory and Representativeness.," *Psychological Review*, 2021, *128* (1), 71.

_ , Nicola Gennaioli, and Andrei Shleifer, "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 2020, *135* (3), 1399–1442.

_ , _ , Yueran Ma, and Andrei Shleifer, "Overreaction in macroeconomic expectations," *American Economic Review*, 2020, *110* (9), 2748–82.

Bower, Gordon H and Michal C Clark, "Narrative stories as mediators for serial learning," *Psychonomic Science*, 1969, *14* (4), 181–182.

Brewer, William F and James C Treyens, "Role of schemata in memory for places," *Cognitive Psychology*, 1981, *13* (2), 207–230.

Bruner, Jerome, "Actual Minds, Possible Worlds (Jerusalem-Harvard Lectures)," 1987.

Bursztyn, Leonardo, Aakaash Rao, Christopher P Roth, and David H Yanagizawa-Drott, "Opinions as Facts," *Review of Economic Studies*, 2023, *90* (4), 1832–1864.

_ , Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth, "Justifying dissent," *The Quarterly Journal of Economics*, 2023, *138* (3), 1403–1451.

Charles, Constantin, "Memory and Trading," *Working Paper*, 2022.

Conlon, John J and Dev Patel, "What Jobs Come to Mind? Stereotypes about Fields of Study," *Working Paper*, 2022.

Danz, David, Lise Vesterlund, and Alistair Wilson, "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 2022, *112* (9), 2851–2883.

Eliaz, Kfir and Ran Spiegler, "A Model of Competing Narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

Enke, Benjamin, "What You See Is All There Is," *Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

_ and Florian Zimmermann, "Correlation Neglect in Belief Formation," *Review of Economic Studies*, 2019, *86* (1), 313–332.

_ , Frederik Schwerter, and Florian Zimmermann, "Associative Memory, Beliefs and Market Interactions," *Working Paper*, 2023.

Foer, Joshua, *Moonwalking with Einstein: The Art and Science of Remembering Everything*, Penguin, 2012.

Fryer, Bronwyn, "Storytelling That Moves People," *Harvard Business Review*, 2003.

Gennaioli, Nicola and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, 2010, *125* (4), 1399–1433.

Graeber, Thomas, "Inattentive inference," *Journal of the European Economic Association*, 2023, *21* (2), 560–592.

_ , **Christopher Roth, and Constantin Schesch**, "Contagious Beliefs," 2023.

_ , **Shakked Noy, and Christopher Roth**, "Lost in Transmission," 2023.

Hartzmark, Samuel, Samuel Hirshman, and Alex Imas, "Ownership, Learning, and Beliefs," *Quarterly Journal of Economics*, 2021, *136* (3), 1665–1717.

Heath, Chip and Dan Heath, *Made to Stick: Why Some Ideas Survive and Others Die*, Random House, 2007.

Hossain, Tanjim and Ryo Okui, "The Binarized Scoring Rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

Jiang, Zhengyang, Hongqi Liu, Cameron Peng, and Hongjun Yan, "Investor Memory and Biased Beliefs: Evidence from the Field," *Working Paper*, 2022.

Kahana, Michael Jacob, *Foundations of Human Memory*, OUP USA, 2012.

Kendall, Chad W and Constantin Charles, "Causal Narratives," *Working Paper*, 2022.

Kensinger, Elizabeth A and Daniel L Schacter, "Memory and Emotion," 2008.

Kwon, Spencer Yongwook and Johnny Tang, "Extreme Events and Overreaction to News," *Working Paper*, 2023.

Link, Sebastian, Andreas Peichl, Christopher Roth, and Johannes Wohlfart, "Attention to the Macroeconomy," *Working Paper*, 2023.

Malmendier, Ulrike and Laura Veldkamp, "Information resonance," Technical Report, Working Paper 2022.

_ **and Stefan Nagel**, "Depression babies: Do macroeconomic experiences affect risk taking?," *The quarterly journal of economics*, 2011, *126* (1), 373–416.

_ **and** _ , "Learning from inflation experiences," *The Quarterly Journal of Economics*, 2016, *131* (1), 53–87.

Mandler, Jean M., *Stories, Scripts, and Scenes: Aspects of Schema Theory*, Erlbaum, 1984.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review*, 2019, *109* (10), 3437–74.

McAdams, Dan P, "Narrative Identity," *Handbook of Identity Theory and Research*, 2011, pp. 99–115.

McRae, Ken and Michael Jones, *14 Semantic Memory*, Vol. 206, Oxford University Press Oxford, 2013.

Michalopoulos, Stelios and Melanie Meng Xue, "Folklore," *Quarterly Journal of Economics*, 2021, *136* (4), 1993–2046.

Monarth, Harrison, "The Irresistible Power of Storytelling as a Strategic Business Tool," *Harvard Business Review*, 2014.

Morag, Dor and George Loewenstein, "Narratives and Valuations," *Working Paper*, 2021.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, 2022, *54* (4), 1643–1662.

Schacter, Daniel L, *Searching for Memory: The Brain, The Mind, And The Past*, Basic books, 2008.

Schank, Roger C. and Robert P. Abelson, *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*, Erlbaum, 1977.

Shepard, Roger N, "Recognition memory for words, sentences, and pictures," *Journal of verbal Learning and verbal Behavior*, 1967, *6* (1), 156–163.

__ **and Lynn A Cooper**, *Mental images and their transformations.*, The MIT Press, 1986.

Shiller, Robert J, "Narrative Economics," *American Economic Review*, 2017, *107* (4), 967–1004.

__ , *Narrative Economics*, Princeton University Press, 2020.

Standing, Lionel, "Learning 10000 pictures," *Quarterly Journal of Experimental Psychology*, 1973, *25* (2), 207–222.

Online Appendix: Stories, Statistics, and Memory

Thomas Graeber Christopher Roth Florian Zimmermann

Summary of the Online Appendix

Section A provides an overview of various additional robustness experiments. Section B presents mechanism evidence from earlier versions of the paper. Section C displays additional tables and figures. Section D provides an overview of the different stories we used in the experiments. Section E illustrates details on the implementation of the randomization. Appendix F provides details on our hand-coding scheme. Appendix G explains how one can compute the Bayesian benchmarks for our setting and gives intuitions. Finally, Appendix H provides proofs for our theoretical results.

A Robustness: Additional Results

A.1 Uninformative Stories

Uninformative qualitative content. The design is identical to our baseline except that we explicitly tell respondents that the anecdotal details do not carry any information above and beyond the number of positive reviews among the collection of randomly drawn reviews. In particular, they receive the following prompt:

> There is a possibility that you will also receive additional anecdotal details about a reviewer and their experience with the product. Note that these additional, anecdotal details do not carry any information above and beyond the number of positive reviews among the collection of randomly drawn reviews. In other words, what matters for your guess is the number of positive reviews among the collection of randomly drawn reviews.

To ensure that our respondents actually internalize this information they need to pass the following comprehension check:

> Which of the following two statements is true?
>
> - What matters for my guess is the number of positive reviews among the collection of randomly drawn reviews.
> - What matters for my guess are only the anecdotal details about the reviewer and their experience with the product cannot help me make a better guess.

Willingness to pay elicitation. We also elicit a hypothetical willingness to pay for each product after the respective belief elicitation, both in the initial and follow-up survey. Respondents are provided with the typical price of the respective products with average reviews as an anchor. For example, in the case of the bicycle they receive the following instructions:

> Assuming you were in need of a bicycle, how much would you be willing to pay for this bike?
>
> To provide a reference, the typical price of a bicycle with average reviews is $600.

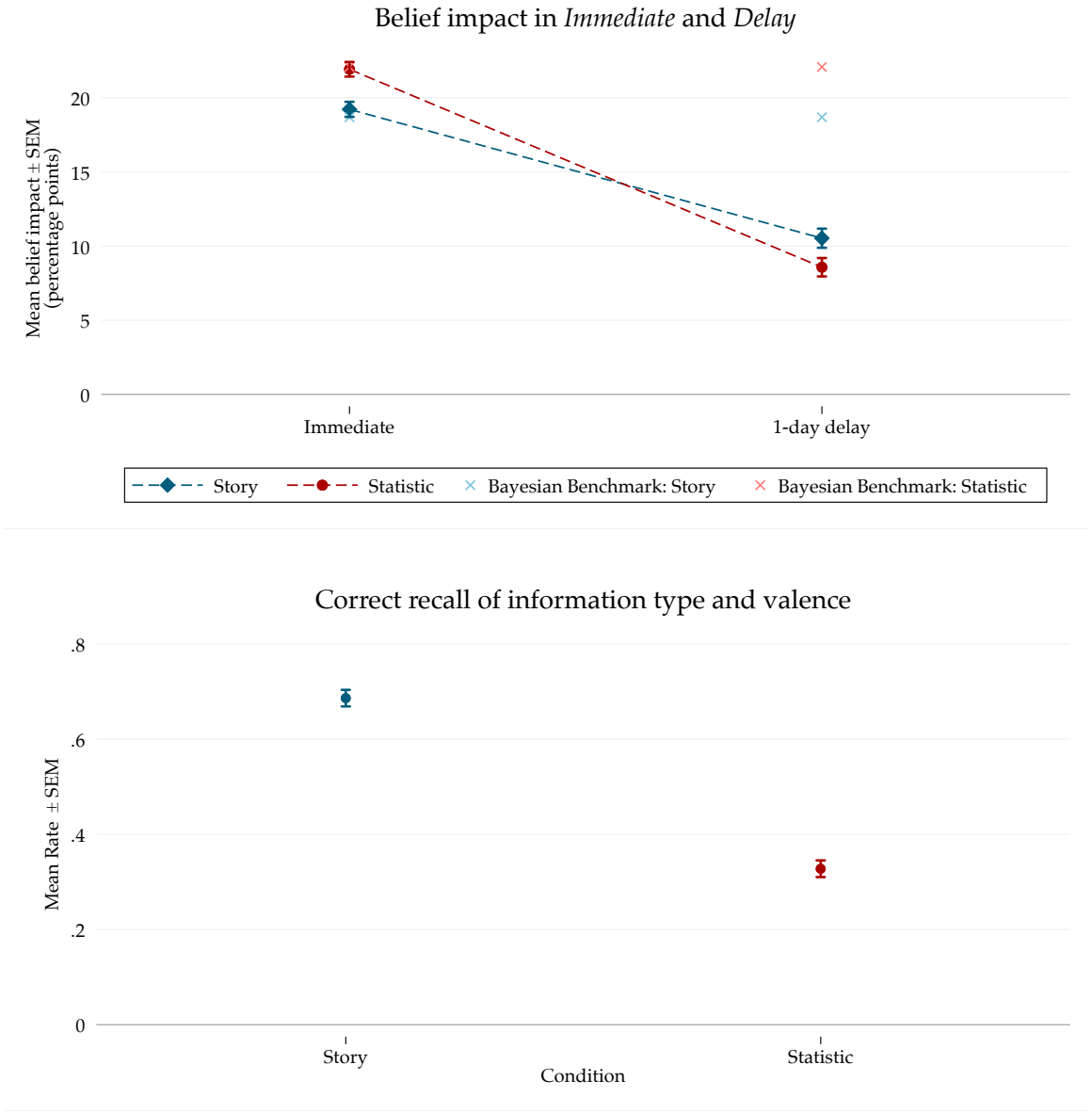


Figure A.1: The story-statistic gap in Robustness experiment 1: Uninformative Qualitative Content (714 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The red markers illustrate belief impact and recall for statistics, while the blue markers illustrate belief impact and recall for stories. The light blue markers illustrate the average Bayesian benchmark for stories (18.70 p.p.), while the light red markers diplays the average Bayesian benchmark for statistics (22.07 p.p.). Whiskers indicate one standard error of the mean.

Table A.1: The story-statistics gap in memory — uninformative stories

	Dependent variable:			
	Belief Impact			Correct Recall
Sample:	Immediate	Delay	Pooled	Pooled
	(1)	(2)	(3)	(4)
Story	-2.70***	1.96*	-2.70***	0.36***
	(0.90)	(1.15)	(0.74)	(0.02)
Delay			-13.3***	
			(0.78)	
Story × Delay			4.66***	
			(0.99)	
Control Mean	21.91	8.57	21.91	0.33
Observations	1428	1428	2856	1428
R^2	0.60	0.59	0.47	0.13

Notes. This Table uses responses from the *Story* and *Statistic* condition in Robustness experiment 1. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Story* takes value 1 for respondents who received a story for a given product, and zero otherwise. *Statistic* takes value 1 for respondents who received a statistic for a given product, and zero otherwise. All columns include respondents who received consistent stories. Column (3) pools *Immediate* and *Delay*. All columns display results on belief impact. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: Impact on willingness to pay

	Dependent variable: Standardized WTP Impact	
Sample:	Story vs. No Info (1)	Statistic vs. No Info (2)
Constant	-0.53***	-0.53***
	(0.04)	(0.04)
Story	0.81***	
	(0.05)	
Statistic		0.73***
		(0.05)
Delay	-0.07**	-0.07**
	(0.03)	(0.03)
Story × Delay	0.11***	
	(0.03)	
Statistic × Delay		0.09**
		(0.04)
Observations	2766	2816
R^2	0.18	0.15

Notes. Column (1) of this table uses responses from the *Story* and *No Information* conditions, while column (2) uses the *Statistic* and *No Information* conditions. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Story* takes value 1 for respondents who received a story for a given product, and zero otherwise. *Statistic* takes value 1 for respondents who received a statistic for a given product, and zero otherwise. All columns include respondents who received consistent stories. All columns display results on standardized WTP impact. Standardized WTP impact is the absolute distance between the stated WTP and the provided anchor, winsorized at 1st and 99th percentile and standardized separately for the six categories (Bicycle, Restaurant, Video game) × (*Immediate*, *Delay*). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.2 Valence of Story Content

To examine the importance of the valence of the story content, our baseline experiment cross-randomized whether the qualitative information in the stories was (i) consistently positive or negative in line with the review rating, (ii) of mixed valence, or (iii) neutral (see Appendix D for all stories). Mixed valence stories mention both positive and negative aspects of the scenario. Neutral valence stories, on the other hand, do not contain evaluations, but describe the experience in the scenario without judgment.

Figure A.2 shows that the valence of story content has minor but significant effects.[1] Average correct recall is 62.15 percent in the consistent story condition compared to 59.58 and 51.20 percent in the mixed and neutral stories treatments, respectively. These levels of recall are substantially higher compared to 26.90 percent for statistics, indicating that the story-statistic gap is robust to variations in the valence of the story content. The patterns for belief impact are consistent with the recall evidence. While belief impact in *Immediate* does indeed depend on the valence of the qualitative information, these differences are strongly attenuated in *Delay*.

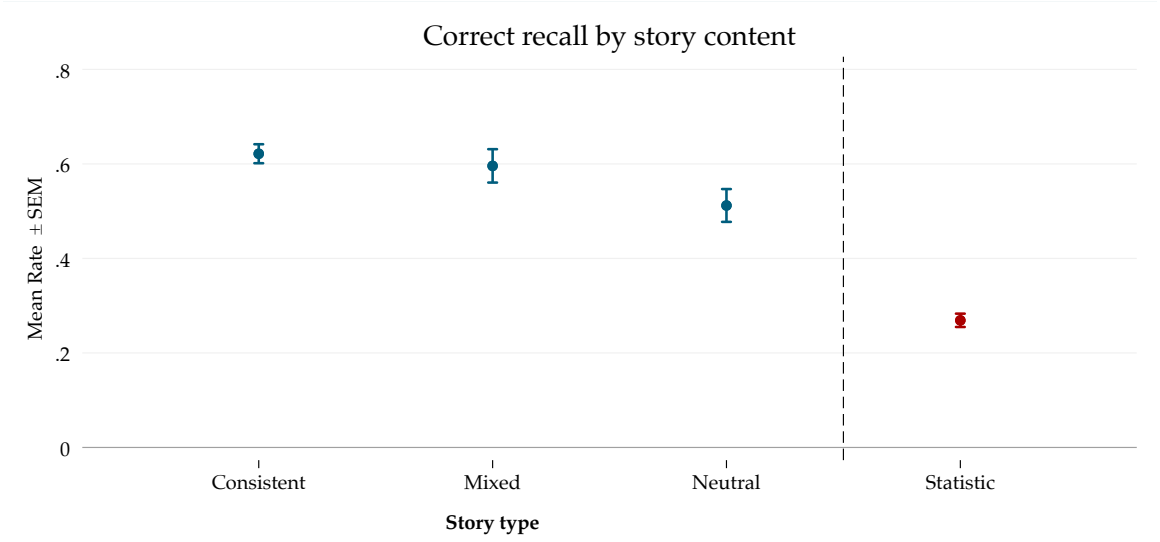


Figure A.2: Correct recall of type and direction by story type in the baseline experiment (984 respondents). The figure shows the fraction of correct recall of the type and direction of information received in the baseline survey in *Delay* for respondents in the *Story* condition (blue) and *Statistic* condition (red). Consistent refers to stories with qualitative features whose direction was fully consistent with the direction of the review. Mixed refers to stories with qualitative features whose direction is mixed. Neutral refers to stories with qualitative features whose direction is neutral. Whiskers indicate one standard error of the mean.

[1] Since we expected the valence manipulation to have potentially strong effects on immediate updating, we pre-registered using recall performance as our main outcome measure.

A.3 Heterogeneity by Positive Versus Negative Reviews

We test for potential heterogeneity in belief impact and correct recall between positive and negative reviews. Figure 2 in the main text already illustrates that there is a pronounced story-statistic gap for both positive and negative reviews. In fact, we find no difference in recall performance, whether the reviews are positive or negative (*Story*: $p = 0.328$, *Statistic*: $p = 0.991$). Moreover, there is no heterogeneity in the evolution of belief impact for *Story* by the direction of the quantitative information ($p = 0.860$) . However, we observe that positive statistics affect beliefs more persistently than negative statistics ($p < 0.001$).

A.4 Robustness to Different Non-Target Information

We exogenously manipulate the type of information for the two non-target scenarios. Respondents either received two statistics for the non-target scenarios, two stories or twice no information. In addition, in contrast to the baseline design, we fully randomize the direction of the information provided for each scenario. In the follow-up survey, we elicit beliefs exactly as in the baseline experiment, presented in Section 3.1.

Sample. We recruited 2,250 respondents for the baseline survey. 2048 respondents qualified for the follow-up survey. 1,613 respondents completed the one-day follow-up survey. After the pre-specified sample restrictions, our final sample consists of 1,548 respondents, corresponding to a 76% completion rate.[2]

Results. Figure A.3 summarizes our results. The left-hand panel shows the changes in belief impact between immediate and delay for the target story and target statistic across the three different conditions. The right panel analogously displays the rate of correct recall across the three conditions separately for the story and statistic target.

We make three observations: First, there is a robust story-statistic gap across all conditions. The story-statistic gap has a similar magnitude irrespective of the number and type of non-target information. This is visible across both our beliefs data and the incentivized structured recall elicitation.[3] Second, we observe small effects at best of the number of decoy information. This suggests that memory load per se has muted effects on belief impact in this setting. Third, we do not observe significant effects of the type of decoy information on the size of the story-statistic gap. Jointly these results imply that the story-statistic gap is robust to basic features of the decoys and that – in a setting with only three scenarios – the type and number of decoys is not a key driver of the decay of belief impact.

Figure A.4 shows how belief impact and recall of stories vary depending on the direction of decoy information. Compared to the statistics benchmark, we again find a robust and sizable story-statistic gap across decoys of different direction. We further find that decoy direction has a small but directionally plausible effect on the size of the gap: when decoy information has the same direction as the target information, both recall and delayed belief impact is larger than when the decoy information is mixed or of opposite sign.

[2]The completion rate to the follow-up survey does not differ significantly across treatment groups ($p = 0.60$).

[3]Results from our structured recall task are very similar to results from the free recall task, providing a validation of the latter.

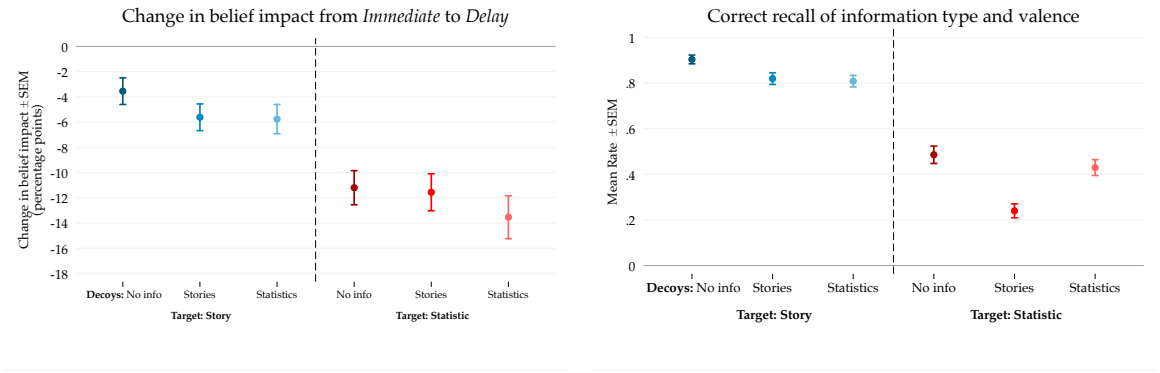


Figure A.3: Belief impact and recall in Robustness Experiment 2: The role of Decoy Information (1,513 respondents). The left panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The dark blue (dark red) markers illustrate change in belief impact and recall for stories (statistics) for the *Decoys: No Info* condition, the blue (red) markers illustrate the change in belief impact and recall for stories (statistics) for the *Decoys: Stories* condition, while the light blue (light red) markers display the change in belief impact and recall for stories (statistics) for the *Decoys: Statistics* condition. Whiskers indicate one standard error of the mean.

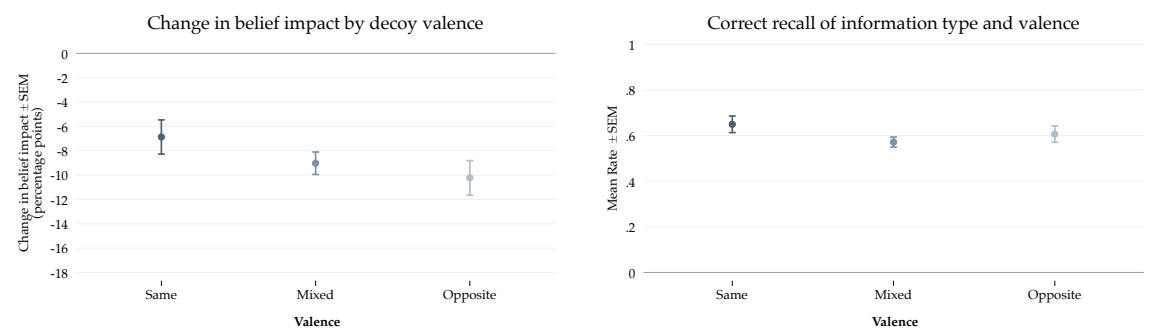


Figure A.4: Belief impact and recall in Robustness Experiment 2: The role of Decoy Information (1,513 respondents). The top panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The right panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The dark gray markers illustrate change in belief impact and recall for targets when decoys have the target's direction, the gray markers illustrate change in belief impact and recall for targets when decoys have mixed direction, while the light gray markers display the change in belief impact and recall for targets when decoys have the target's opposite direction. Whiskers indicate one standard error of the mean.

A.5 The Number of Product Scenarios

In this section we examine the robustness of our findings to varying the number of products. We examine how the size of the story-statistic gap varies depending on whether there are one, three or six products.

Design. The design broadly follows the structure of the main experiment. The key difference is that we vary, between subjects, whether there are one, three or six product scenarios. In the *1-product* treatment, there is a single scenario and participants only received one piece of information, either a story or a statistic. Identical to the baseline experiment, participants in the *3-product* treatment see three scenarios and receive two pieces of information, one story, one statistic and once no information. In the *6-product* treatment, participants see six scenarios overall and also receive two pieces of information (one story and one statistic), as well as four times no information. This means that the comparison between the *3-product* and *6-product* design allows us to cleanly study the effects of the number of product scenarios, while holding the total pieces of information constant; in other words respondents in both the *6-product* and *3-product* treatments receive one statistic and one story.[4]

To keep incentives exactly constant between the different conditions, participants in all treatments complete a total of six payoff-relevant tasks in both *Immediate* and *Delay*: the additional filler tasks are incentivized dot estimation tasks. Respondents in the 1-product treatment arm complete 5 dot estimation tasks, while respondents in the 3-product treatment arm complete 3 dot estimation tasks, and respondents in the 6-product treatment only face product-related tasks.

Sample. We recruited 1500 respondents. 1404 respondents qualified for the follow-up survey. After the pre-specified sample restrictions, our final sample consists of 1018 respondents, corresponding to a completion rate of 73 percent.[5]

Results. Figure A.5 and Table A.3 illustrate changes in belief impact between *Immediate* and *Delay* as well as recall for stories and statistics across the different number of product scenarios. The top panel depicts the change in belief impact between *Immediate* and *Delay* across the three treatment arms, separately for stories and statistics. We find that, overall, the change in belief impact tends to become more pronounced as we increase the number of product scenarios. This effect is relatively small for stories. In fact, the *6-product* treatment does not lead to a more pronounced decay of belief impact than the *3-product* and *1-product* versions. At the same time, the effect of more scenarios on the decay of belief impact is quantitatively large for statistics. As a consequence, and in line with our model, the story-statistic gap widens with the number of product scenarios.[6]

[4]The comparison between the *1-product* and *3-product* condition jointly identifies the effects of increasing the total number of products and increasing the pieces of information.

[5]The completion rate to the follow-up survey does not differ significantly across treatment groups ($p = 0.37$).

[6]The story-statistic gap in belief impact is close to zero for the 1-product scenario.

This pattern is strongly supported by the recall data, see the bottom panel of Figure A.5. Recall accuracy of statistics drastically decreases as we move from 1 to 3 to 6 scenarios, while recall accuracy of stories remains comparably stable.

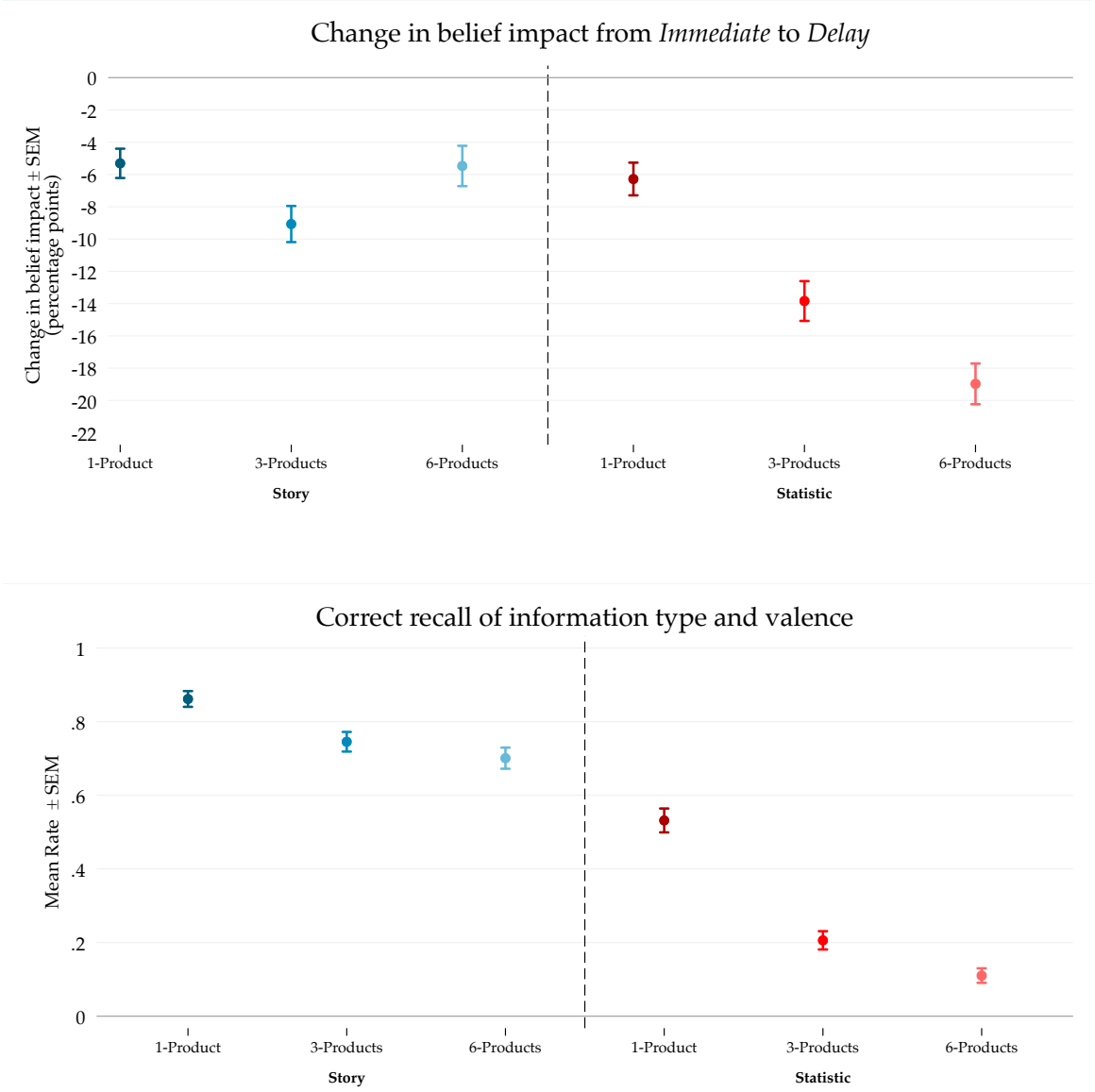


Figure A.5: Change in belief impact and recall in Robustness Experiment 3: Number of product scenarios (1,018 respondents). The top panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The dark blue markers illustrate change in belief impact and recall for the *1-product* condition, the blue markers illustrate the change in belief impact and recall for the *3-product* condition, while the light blue markers display the change in belief impact and recall for the *6-product* condition. Whiskers indicate one standard error of the mean.

Table A.3: The story-statistic gap by number of products

	Dependent variable:			
	Belief Impact		Combined Recall	
Sample:	Story (1)	Stat (2)	Story (3)	Stat (4)
1-Product	-1.02	2.26	0.12***	0.33***
	(1.39)	(1.44)	(0.03)	(0.04)
Delay × 1-Product	3.76***	7.52***		
	(1.44)	(1.59)		
6-Products	-1.44	2.76**	-0.045	-0.096***
	(1.49)	(1.38)	(0.04)	(0.03)
Delay × 6-Products	3.60**	-5.13***		
	(1.68)	(1.76)		
Delay	-9.07***	-13.8***		
	(1.12)	(1.23)		
Control Mean	18.48	18.51	0.75	0.21
Observations	1562	1515	781	758
R^2	0.04	0.19	0.03	0.16

Notes. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *1-Product* is an indicator taking value 1 if the respondent receives one product scenario and 0 else. *6-Products* is an indicator taking value 1 if the respondent receives six product scenarios and 0 else. Columns (1) and (3) include respondents who received stories, while column (2) and (4) include respondents who received statistics. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Columns (3) and (4) display the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.6 Robustness to Elicitation Format

It is conceivable that the question format of the belief elicitation affects the story-statistic gap, because the specific wording of the question might favor the recall of stories over statistics. In an additional experiment, we independently manipulated the question format as well as the display format of the statistic. First, the *Likelihood Format* treatment elicited beliefs as before – about the likelihood that a randomly chosen review is positive – and thus exactly corresponded to our main experiment. In the *Fraction Format* condition, by contrast, we elicited beliefs about the percentage of positive reviews in the overall population of reviews of the product. Second, we randomized whether the statistical information itself was expressed in terms of an absolute number of positive reviews in a subsample – as in our main study – (*Statistic Number Display*) or in terms of a percentage of positive reviews in a subsample (*Statistic Percent Display*). Figure A.6 shows that the fraction question format has a positive, albeit small effect on delayed belief impact and recall. Moreover, displaying statistical information as a percentage instead of an absolute number does not have significant effects on belief impact and recall. We also do not observe a significant interaction effect between the question format and the display format of statistical information. Taken together, this evidence highlights that the story-statistic gap in memory is robust to the exact question format used.

Given that the way the statistical information is presented should affect the computational complexity of calculating immediate beliefs, our evidence provides suggestive evidence that computational complexity and the associated cognitive load do not seem to play a quantitatively important role.

Table A.4: Question format: belief impact and recall

	Dependent variable:			
	Belief Impact		Combined Recall	
Sample:	Story (1)	Stat (2)	Story (3)	Stat (4)
Similar Format	1.95*	0.53	0.0094	0.085**
	(1.10)	(1.27)	(0.03)	(0.04)
Delay × Similar Format	-0.63	1.45		
	(1.31)	(1.88)		
Statistic Similar		1.98		0.019
		(1.30)		(0.04)
Delay × Statistic Similar		-0.15		
		(1.84)		
Statistic Similar × Similar Format		-1.68		-0.064
		(1.78)		(0.06)
Delay × Statistic Similar × Similar Format		-1.28		
		(2.60)		
Delay	-8.19***	-14.7***		
	(0.87)	(1.33)		
Control Mean	18.50	20.63	0.73	0.19
Observations	1718	1718	859	859
R^2	0.06	0.19	0.00	0.01

Notes. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Similar Format* takes value 1 for respondents whose beliefs were elicited in percent. *Statistic Similar* is an indicator taking value 1 for respondents who received statistics in a percentage format. Columns (1) and (3) include respondents who received stories. Columns (2) and (4) include respondents who received statistics. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Columns (3) and (4) display the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

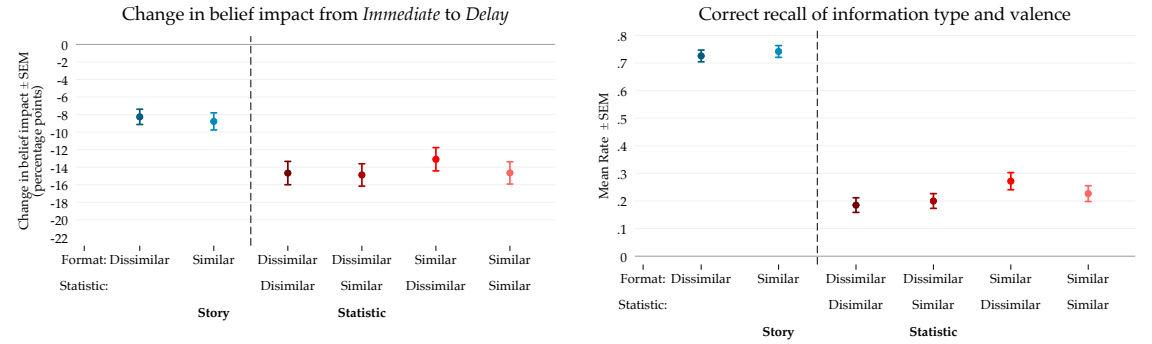


Figure A.6: Belief impact and recall in Robustness Experiment 4: Question Format and statistic display (959 respondents). The left panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The right panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The dark blue markers illustrate change in belief impact and recall for the *Dissimilar Format* condition for stories, the light blue markers illustrate change in belief impact and recall for the *Similar Format* condition for stories, while the most dark red for the *Dissimilar Format / Statistic Dissimilar* condition, the dark red for the *Dissimilar Format / Statistic Similar* condition, the red for the *Similar Format / Statistic Dissimilar* condition and the light red for the *Similar Format / Statistic Similar* condition. Whiskers indicate one standard error of the mean.

A.7 The Role of Qualitative Associations

Design. To causally examine the role of adding qualitative features *while holding the amount of information content provided constant*, we prompt respondents to imagine a typical review for the statistic or for a single review they learn about. This treatment does not provide any objective information, qualitative or quantitative, allowing us to identify the distinct effect of associating obviously fictional qualitative features with a piece of information in memory.

We implement four conditions. In *Baseline*, we replicate our main design. The *StatisticPrompt* condition is identical to *Baseline*, except that respondents that receive the statistic are prompted "to imagine how a typical review based on the provided information would look like."

To examine the role of associations for single reviews that do not contain any qualitative features, we design two additional treatments. The *NoStory* condition is identical to *Baseline*, except that instead of a story, respondents receive information about a single review without any qualitative information. The *NoStoryPrompt* condition is identical to *NoStory* except that respondents that received information about a single review are asked to imagine what the review might look like, similar to *StatisticPrompt*. The rationale behind these two conditions is to examine what happens when the story provided in the *Story* condition of our main experiment is stripped of its actual content and then replaced by an endogenously generated one. The prompt in turn may push people to

retrieve personal experiences that they have made with similar products in the past.

We use a structured recall task. We ask respondents to indicate whether they (i) received information about a single review, including some additional anecdotal details about the reviewer and their experience with the product, (ii) multiple reviews, (iii) no information or (iv) don't know. Unless respondents indicate that they did not receive any information about this product, we additionally ask them to indicate whether the information they received was positive or negative. Respondents are told that if they correctly recall the information they received, they will receive an additional bonus of $5. To circumvent hedging motives, either beliefs or recall were randomly selected for payment, and one question was randomly chosen to determine the bonus.

Sample and pre-registration. 1,500 respondents completed wave 1 of our experiment, with 1,442 qualifying for wave 2. Of those, 703 respondents actually completed wave 2. 666 of the final set of respondents satisfied our inclusion criteria, corresponding to a completion rate of 46 percent.[7]

Prediction. The decay of belief impact and forgetting is lower in the *Prompt* conditions than in the *No Prompt* conditions.

Results. We start by examining whether the prompt intervention was effective in actually inducing participants to imagine reviews and to write them down. The median (mean) number of words participants wrote to describe an imaginary typical review was 22 (23). The text responses indicate that the vast majority of participants made a significant effort to describe a review, such as in the following excerpt from a response in the *NoStoryPrompt* condition about a negative videogame review:

> *The gameplay was sub-par and glitched randomly. The graphics compared the trailer to the actual gameplay were very different giving the impression that the gameplay will have 3D style graphics while in reality, it had very old-school-style graphics [...].*

For ease of exposition, Figure A.8 pools respondents in *NoStoryPrompt* and *StatisticPrompt*, as well as the *NoStory* and *Baseline* conditions.[8] The top panel of Figure A.8 shows results on belief impact, while the bottom panel displays results on recall.

[7]The completion rate to the follow-up survey does not differ significantly across treatment groups ($p = 0.90$). The somewhat lower completion rate compared to the baseline experiment can be explained by the fact that part of the experiment took place on the weekend.

[8]Table A.5 shows results separately for all 4 conditions and confirms that the disaggregated results are similar.

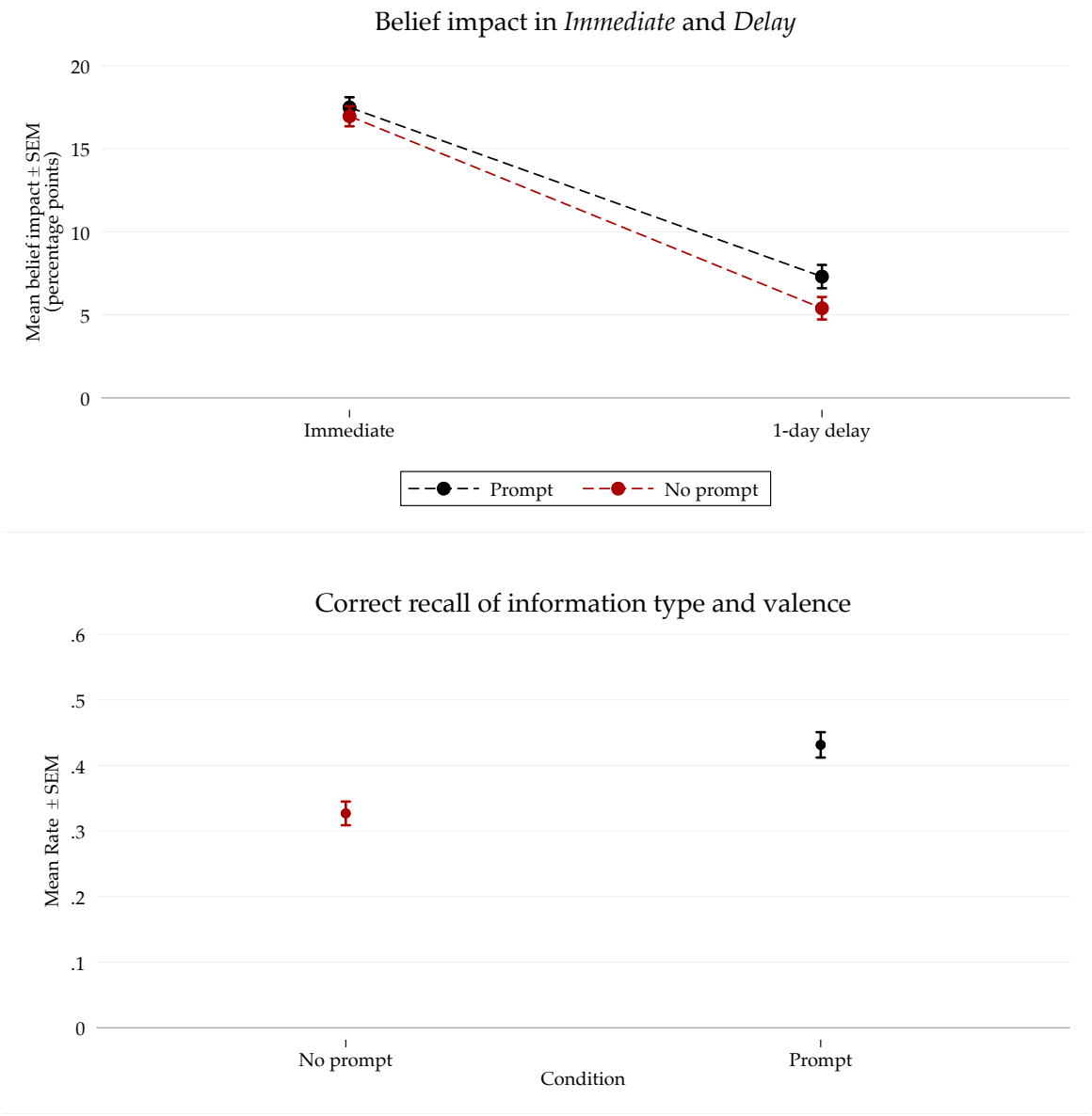


Figure A.7: Belief impact and recall in Robustness Experiment 5 (666 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The black markers illustrate belief impact and recall for *Prompt*, while the red markers illustrate belief impact and recall for *NoPrompt*. Whiskers indicate one standard error of the mean.

Starting with belief impact, we find that, reassuringly, beliefs in *Immediate* are not meaningfully different across the *Prompt* and the *NoPrompt* conditions. Yet, in *Delay*, average belief impact for respondents in the *Prompt* conditions is 7.30 p.p. (s.e. 0.70) compared to only 5.40 p.p. (s.e. 0.68) in *NoPrompt*. This treatment difference in *Delay* is statistically significant ($p < 0.01$). Column (1) of Table A.5 reveals that the difference-in-differences (difference in slopes) is also statistically significant ($p < 0.05$).

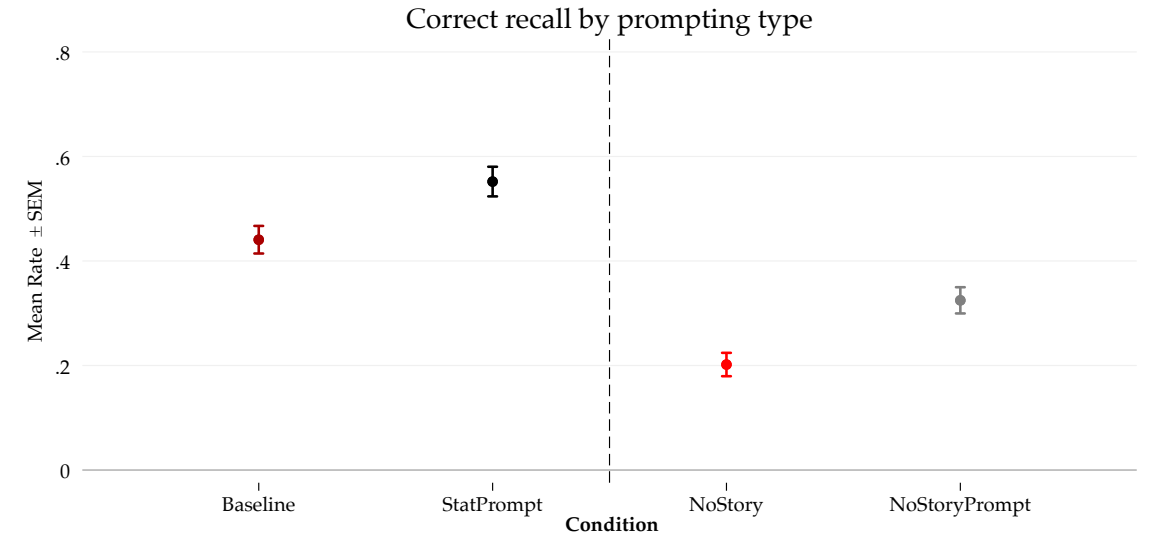


Figure A.8: Belief impact and recall in Robustness Experiment 5: The role of associations (666 respondents). The panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The dark red markers illustrate belief impact and recall for *Baseline*, while the light red markers illustrate belief impact and recall for *NoStory*. The black markers illustrate belief impact and recall for *StatPrompt*, while the gray markers illustrate belief impact and recall for *NoStoryPrompt*. Whiskers indicate one standard error of the mean.

These patterns for *Delay* beliefs are underscored by results on recall. The bottom panel of Figure A.8 shows that recall accuracy is 43.14 percent for respondents in *Prompt*, compared to only 32.69 percent in the conditions without prompt. Table A.5 reveals that these differences are highly statistically significant when comparing respondents in the *StatisticPrompt* and *Baseline* conditions, as well as when comparing respondents in the *NoStoryPrompt* and *NoStory* conditions.

While the effect size from this experiment is smaller than in our baseline evidence, it is worth bearing in mind that approximately 17% of respondents did not fully engage with the prompt. Consistent with the idea that engagement with the prompt matters, respondents with above median text length in the prompt are 11 percentage points more likely to correctly recall the information than those respondents with a below median text length.

Recall of binary quantitative information. One result that emerges from this experiment is that in the absence of a prompt to encode additional qualitative information, people perform similarly at recalling information about a binary variable as they do at recalling a statistic. Specifically, Table A.5 reveals that correct recall among respondents in the *NoStoryPrompt* condition is 16 percent and thus, if anything, lower compared to correct recall of statistical information in the *Baseline* condition (22 percent).

Table A.5: Prompting Experiment: belief impact and recall

	Dependent variable:					
	Belief Impact			Combined Recall		
Sample:	Pooled (1)	Stat (2)	NoStory (3)	Pooled (4)	Stat (5)	NoStory (6)
Delay	-11.5***	-14.7***	-7.95***			
	(0.97)	(1.31)	(1.39)			
Prompt	-0.97	-1.47	1.00	0.20***	0.14***	0.26***
	(1.19)	(1.54)	(1.50)	(0.03)	(0.05)	(0.05)
Delay × Prompt	3.35**	4.22**	1.90			
	(1.34)	(1.93)	(1.83)			
Control Mean	14.47	21.57	6.66	0.19	0.22	0.16
Observations	1332	662	670	1332	662	670
R^2	0.09	0.15	0.06	0.05	0.02	0.08

Notes. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Prompt* is an indicator taking value 1 for respondents who were prompted to imagine a typical review when provided with statistical information. All columns pool *Immediate* and *Delay*. Columns (1) and (4) include all respondents. Column (2) and (4) include respondents who received statistics. Columns (3) and (6) include observations who received information on a single review. Columns (1) to (3) display results on belief impact. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Columns (4) to (6) display the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Previous Mechanism Evidence

B.1 Cue Similarity - Details

We conducted an additional experiment that varied the similarity between different scenario names (Restaurant A, Restaurant B, Restaurant C). This evidence shows that increasing cue similarity decreases forgetting significantly. Because this study design is not directly connected to the framework of Section 2, the corresponding evidence is presented in the Appendix.

Design. Our design varied the similarity of cues, holding everything else constant. The basic set-up follows our main experiment. In *Baseline*, the three cues were Restaurant A, Bicycle and Videogame, with Restaurant always being the target cue in our analysis. Participants either received a story or a statistic in the restaurant scenario. In *Cue Similarity*, we kept everything identical to *Baseline*, including the target cue Restaurant A, but changed the labels of the decoy cues to Restaurant B and Restaurant C. In our

analysis, as pre-registered, we compare belief impact and recall between the *Baseline* and *Cue Similarity*, separately for respondents who received a story and a statistic.

Sample and pre-registration. We recruited 1,150 respondents, of which 999 were eligible for the followup. Out of those, 599 respondents completed the follow-up survey. After the pre-specified sample restrictions, our final sample consists of 583 respondents, corresponding to a completion rate of 59 percent.[9] The pre-registration for this experiment is available at `https://aspredicted.org/h2fr3.pdf`.

Prediction. The decay of belief impact and forgetting of both stories and statistics are more pronounced in *Cue Similarity* than *Baseline*.

Results. Panel A of Figure A.9 displays changes in belief impact between *Immediate* and *Delay* for both treatments. The figure reveals that the change in belief impact is substantially larger in the cue similarity condition. This holds true both when the target is a story and when the target is a statistic (though the effect is less pronounced for statistics, possibly due to already very low levels of delayed belief impact and recall). Panel B of Figure A.9 largely displays the same pattern using our recall data. Table A.6 confirms this result.

[9]The completion rate to the follow-up survey does not differ significantly across treatment groups ($p = 0.53$).

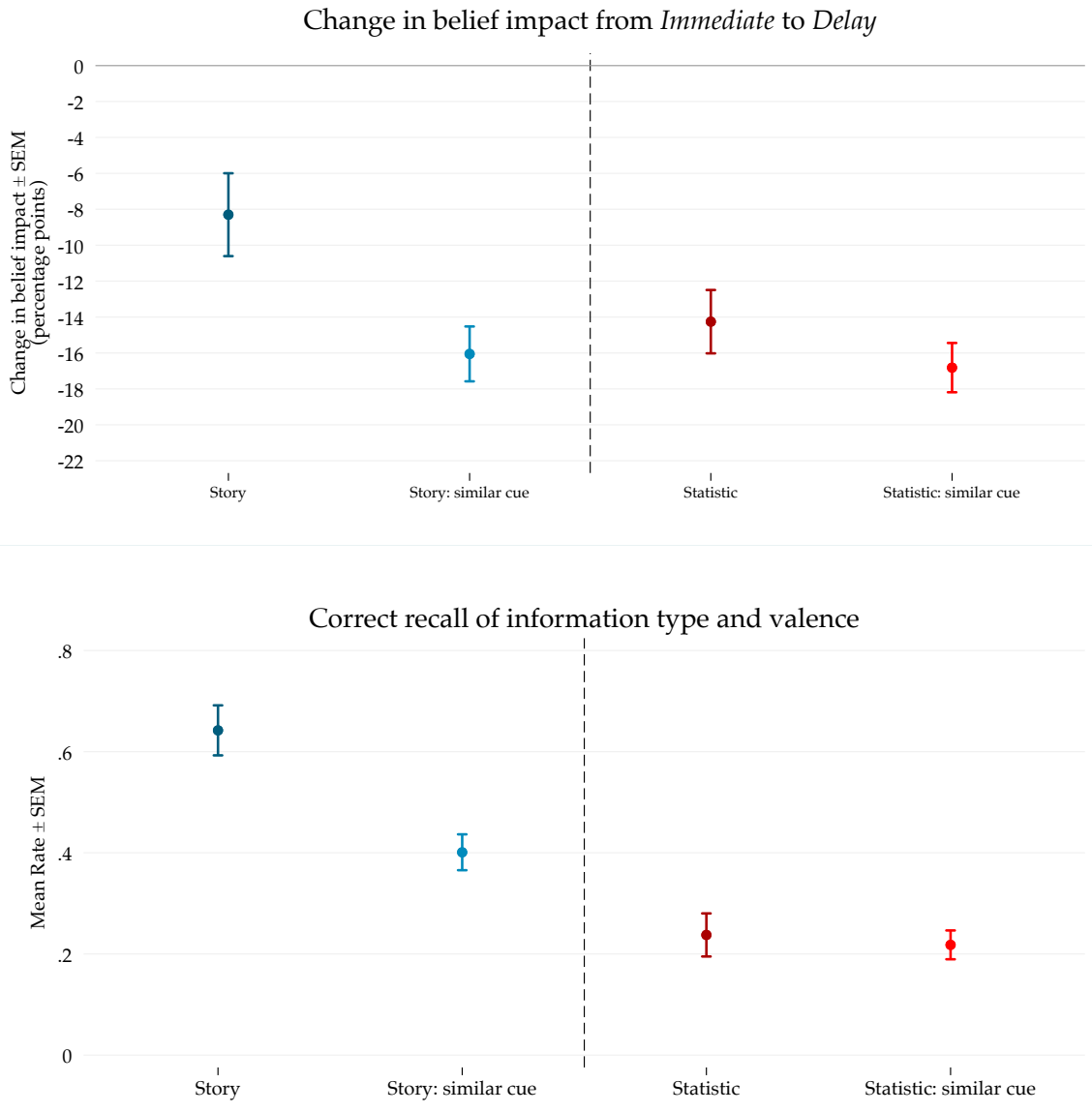


Figure A.9: Change in belief impact and recall in Appendix Mechanism Experiment 1 (1,018 respondents). The left panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The right panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The dark blue markers illustrate change in belief impact and recall for the *Story* condition, while the light blue markers illustrate the change in belief impact and recall for the *Story with Cue Similarity* condition. The dark red markers illustrate change in belief impact and recall for the *Statistic* condition, while the light red markers illustrate the change in belief impact and recall for the *Statistic with Cue Similarity* condition. Whiskers indicate one standard error of the mean.

Table A.6: Cue similarity

	Dependent variable:			
	Belief Impact		Combined Recall	
Sample:	Story	Stat	Story	Stat
	(1)	(2)	(3)	(4)
Similar Cue	0.21	-0.77	-0.24***	-0.020
	(2.13)	(1.68)	(0.06)	(0.05)
Delay × Similar Cue	-7.75***	-2.56		
	(2.77)	(2.23)		
Delay	-8.30***	-14.3***		
	(2.30)	(1.76)		
Control Mean	18.80	21.62	0.64	0.24
Observations	574	624	287	312
R^2	0.14	0.21	0.05	0.00

Notes. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Similar Cue* is an indicator taking value 1 for respondents who received three restaurant scenarios. Columns (1) and (3) include respondents who received stories, while column (2) and (4) include respondents who received statistics. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Columns (3) and (4) display the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.2 Cue-Story Similarity: Previous Evidence

We designed two treatments to study the role of story similarity. The incentives and basic setting are identical to our main experiment. Participants in both conditions learn about three scenarios: a cafe, a restaurant, and a bar. Unlike in our main experiment, respondents receive a story in each of the three scenarios. In *Baseline*, the stories in all scenarios are specific to the product of venue. In the *Cue-Story Similarity*, however, the target story about a bar involved an experience that is entirely unrelated and unspecific to a bar. The objective was to exogenously reduce the similarity between the target story and the target cue, keeping all other design aspects fixed. Appendix D.4 contains the stories that we used.

As specified in the pre-analysis plan (`https://aspredicted.org/v7hh6.pdf`), we focus on the recall data, because the immediate belief impact was likely to be much stronger in *Baseline* than in *Cue-Story Similarity* (as was indeed the case in our data). Column (4) of Table A.7 documents that, while correct recall in the *Baseline* was 47.04 percent (s.e. 0.03), recall in the *Cue-Story Similarity* condition was 40.21 percent (s.e.

0.03) percent. This difference is statistically insignificant ($p = 0.10$). Column (2) of Table A.7 reports results on belief impact. The decay of belief impact points in the opposite direction, i.e., *Cue-Story Similarity* was associated with lower decay of belief impact over time. This result is hard to interpret given different belief impact in immediate, but nevertheless highlights that the overall evidence of this manipulation for cue-target similarity is mixed.

We would like to note that this evidence is now placed in the Appendix for two reasons: First, our manipulation which decreased cue-story similarity at the same time decreased interference, i.e. the similarity of the target story to the non-target stories, as the generic story about the shopping experience should be more dissimilar from the other stories. This confounds the story because cue-story similarity and the interference mechanisms work in opposing directions. Second, the differences in immediate belief updating are relatively large. Hence, it is difficult to interpret the patterns of dynamic belief movements. Taken together, it is difficult to draw clear conclusions from our old evidence. We provide an improved set of experiments in Section 4.1.

Table A.7: (Cue-)story similarity

	Dependent variable:			
	Belief Impact		Combined Recall	
Sample:	Story (1)	Cue-Story (2)	Story (3)	Cue-Story (4)
Story Similarity	2.61**		-0.097**	
	(1.25)		(0.04)	
Delay × Story Similarity	-5.79***			
	(1.78)			
Cue-Story Similarity		-6.21***		-0.068
		(1.21)		(0.04)
Delay × Cue-Story Similarity		4.27***		
		(1.62)		
Delay	-14.2***	-14.2***		
	(1.16)	(1.16)		
Control Mean	18.68	18.68	0.47	0.47
Observations	1136	1136	568	568
R^2	0.21	0.15	0.01	0.00

Notes. This Table shows data from Appendix Mechanism Experiment 2 (849 respondents). *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Story Similarity* takes value 1 for respondents who received similar decoy stories, and zero otherwise. *Cue-Story Similarity* takes value 1 for respondents who received a generic story that was less intrinsically related to the cue compared to the baseline condition. Columns (1) and (3) include respondents who were in the story similarity and baseline condition. Columns (2) and (4) include respondents who were in the cue-story similarity and baseline condition. Columns (1) and (2) display results on belief impact. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Columns (3) and (4) display the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. OLS estimates, standard errors clustered at the respondent level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.3 Story Similarity: Previous Evidence

Below we present our previous evidence on story similarity.

Design. We designed two treatments to study the role of story similarity. The incentives and basic setting are identical to our main experiment. Participants in both conditions learn about three products: a cafe, a restaurant, and a bar. Unlike in our main experiment, respondents receive a story in each of the three scenarios. The target story in both conditions that our analysis focuses on is a positive review about the bar. The stories about the restaurant and the cafe are decoy stories and both featured a negative review. In the *Baseline* condition, the three stories are distinct and specific to each cue. The bar story describes the interior of the bar, the restaurant story focuses on food quality, while the cafe story is concerned with the service quality. In the *Story Similarity* condition, we keep the target story about the bar identical to *Baseline*, but increase the similarity of the two decoy stories to the target story by modifying both the text structure and content. Specifically, in *Story Similarity*, the three products are still a cafe, a restaurant, and a bar, but all stories revolve around the interior design of the respective venues. Thus, our treatments fix the target story and only manipulate the similarity between the two decoy stories and the target story. All other design aspects are identical between the conditions. Appendix D.4 reproduces all stories that we used.

Sample. We recruited 1,150 respondents, of which 1,069 qualified for the follow-up. Respondents were randomized into the two conditions described above. 879 respondents completed the follow-up survey. After the pre-specified sample restrictions, we have a sample size of 849, corresponding to a completion rate of 79 percent.[10]

Results. The top panel of Figure A.10 shows data on the belief impact of the target story in *Immediate* and *Delay*, separately for *High Similarity* and *Baseline*.[11] In line with the model prediction, the slope in belief impact is steeper in *High Similarity* compared to *Baseline*. Delayed belief impact is significantly lower in *High Similarity* than in *Baseline*,

[10]The completion rate to the follow-up survey does not differ significantly across treatment groups ($p = 0.79$).

[11]Forgetting in *Baseline* of this mechanism experiment is higher than in our baseline experiment for potentially three reasons: First, the cues in this experiment are more similar to each other compared to the baseline experiment. Second, the information provided in the baseline condition is more similar to each other compared to the information in the baseline condition given that they all come in a story format. Third, respondents in this mechanism experiment receive three pieces of information rather than only two pieces of information.

even though immediate belief impact is larger in the former condition.[12]

While average delayed belief impact in *High Similarity* is 1.25 p.p. (s.e. 1.17), it is 4.43 p.p. (s.e. 1.09) in *Baseline*. Table A.7 confirms this visual pattern and shows that the difference-in-differences in belief impact (difference in slopes) is statistically significant ($p < 0.01$).

The bottom panel illustrates similar patterns for recall: Among respondents in *Baseline*, 47.04 p.p. (s.e. 0.03) correctly recall the information, compared to only 37.37 p.p. (s.e. 0.03) in *High Similarity*. This difference of 10 p.p. is statistically significant ($p < 0.01$). This effect size is moderate in size and corresponds to 0.20 of standard deviation

[12]Immediate belief impact might be larger in *High Similarity* due to a more pronounced contrast effect when the target story is more similar to the decoy stories. Naturally, higher immediate belief impact works against us finding differences in the *Delay* condition and also does not affect our evidence on recall.

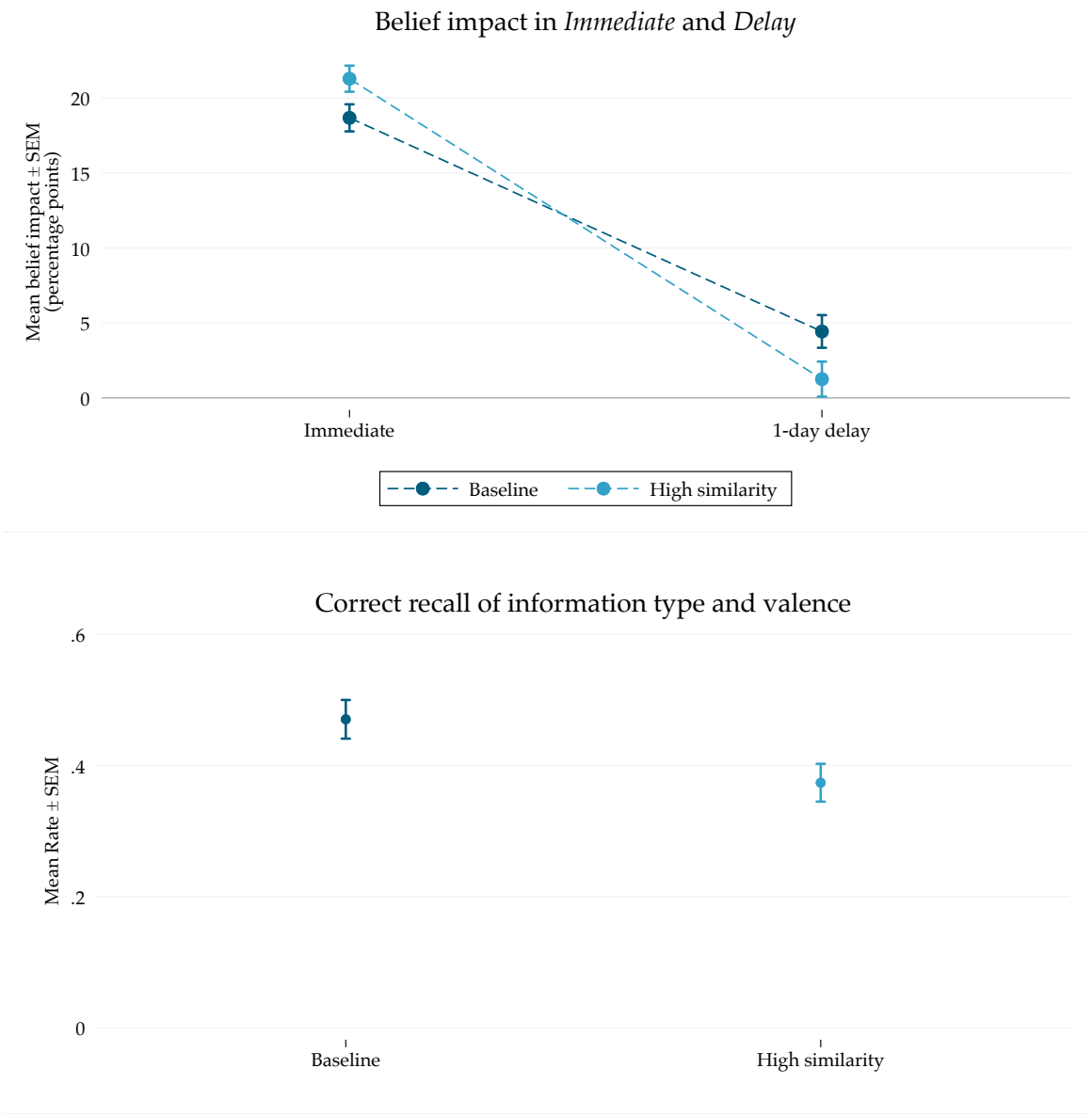


Figure A.10: Belief impact and recall in Appendix Mechanism Experiment 2 (849 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The dark blue markers illustrate belief impact and recall for *Low similarity to non-target information*, while the light blue markers illustrate belief impact and recall for *High similarity to non-target information*. Whiskers indicate one standard error of the mean.

C Additional Tables and Figures

Table A.8: Overview of data collections

Collection	Sample	Baseline Treatments	Additional Treatments	Main outcomes	Link to pre-analysis plan
Baseline experiments					
Baseline Experiment	Prolific (984 respondents)	3 products: story, statistic, no information	For story treatment 3 different types of qualitative features: consistent, neutral, mixed.	Beliefs in immediate and delay; Open-ended recall in delay	https://aspredicted.org/e5mw7.pdf
Statistics with Qualitative Content	Prolific (673 respondents)	3 products: statistic with qualitative content, statistic without qualitative content, no information	None	Beliefs in immediate and delay; Structured incentivized recall in delay	https://aspredicted.org/2RB_H9J
Main Mechanism Experiments					
Mechanism Experiment 1: Cue-Target Similarity	Prolific (627 respondents)	3 products: story, statistic, no information	**High Similarity**: The Italian Restaurant "Napoli". **Low Similarity 1:** An Eatery **Low Similarity 2:** Mr. Jones	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/D21_PW7
Mechanism Experiment 2: Cue-Non-Target Similarity	Prolific (505 respondents)	3 venues: food truck, sports stadium and amusement park	**Low Similarity**: 3 distinct stories. **High similarity**: same story about target product, but now similar stories about other products.	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/4Q6_3YD
Robustness					
Robustness experiment 1: Uninformative Qualitative Content	Prolific (714 respondents)	3 products: story, statistic, no information	None	Beliefs in immediate and delay; Structured incentivized recall in delay	https://aspredicted.org/B49_DB1
Robustness Experiment 2: The role of Decoy Information	Prolific (1,513 respondents)	3 products (1 target and 2 non-target products): Target: Either Story or Statistic	Non-Target products: Either 2 stories, 2 statistics or 2 times no information	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/qy3wq.pdf
Robustness Experiment 3: Number of product scenarios	Prolific (1,018 respondents)	1 product: Statistic or story; 3 products (statistic, story, no info; 6 products: statistic, story and 4 times no info	None	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/as7i7.pdf
Robustness Experiment 4: Question Format and statistic display	Prolific (959 respondents)	3 products: story, statistic, no information	**Likelihood**: elicitation from baseline. **Fraction**: elicitation about the percentage of positive reviews **Statistic number:** number of positive reviews. **Statistic percent**: percentage of positive reviews.	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/ZFF_88V
Robustness Experiment 5: The role of associations	Prolific (666 respondents)	3 products. Decoys: Story and no information; Target varies across treatments	**Baseline**: statistic without prompt; **Prompt**: statistic with prompt; **No story**: Single review without prompt; **No story prompt**: Single review with prompt	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/v9gk7.pdf
Appendix Mechanism Experiments					
Appendix Mechanism Experiment 1: Cue similarity	Prolific (583 respondents)	3 products: story, statistic, no information	**Baseline condition**: Restaurant A, Bicycle, Videogame; **Cue similarity condition**: Restaurant A, Restaurant B and Restaurant C	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/h2fr3.pdf
Appendix Mechanism Experiment 2: Story similarity and Cue-story similarity	Prolific (849 respondents)	3 products (bar, cafe and restaurant) with 3 stories	**Low Similarity**: 3 distinct stories. **High similarity**: same story about target product, but now similar stories about other products. **Cue-story similarity**: Same non-target info as in Low Similarity, but the target story entirely unrelated to the target product.	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/v7hh6.pdf

This Table provides an overview of the different data collections. The sample sizes refer to the final sample of respondents that completed both waves and satisfied the pre-specified inclusion criteria for each of our collections.

Table A.9: The story statistics gap: page time heterogeneity

	Dependent variable:			
	Belief Impact			Recall combined
Sample:	Immediate (1)	Delay (2)	Pooled (3)	Consistent (4)
Story	-2.67**	4.78***	-2.67**	0.35***
	(1.32)	(1.51)	(1.32)	(0.04)
Delay			-14.7***	
			(1.11)	
Story × Delay			7.46***	
			(1.56)	
Slow	0.39	-0.34	0.39	0.088**
	(1.17)	(1.39)	(1.17)	(0.04)
Story × Slow	0.60	3.91*	0.60	-0.026
	(1.80)	(2.13)	(1.80)	(0.05)
Delay × Slow			-0.72	
			(1.55)	
Story × Delay × Slow			3.31	
			(2.23)	
Control Mean	20.44	5.76	20.44	0.23
Observations	1168	1168	2336	1168
R^2	0.01	0.04	0.11	0.13

Notes. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Story* takes value 1 for respondents who received a story for a given product, and zero otherwise. *Slow* is an indicator taking value 1 for respondents whose response time was above the median in their condition. Columns (1), (2), (3) and (4) include respondents who received consistent stories. Column (3) pools *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Columns (4) displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.10: Statistics with Qualitative Content

	Dependent variable:			
	Belief Impact			Correct Recall
Sample:	Immediate (1)	Delay (2)	Pooled (3)	All (4)
Statistic with Context	3.41***	8.41***	3.41***	0.37***
	(0.95)	(1.25)	(0.78)	(0.02)
Delay			-18.8***	
			(0.87)	
Statistic with Context × Delay			5.00***	
			(1.15)	
Control Mean	21.95	3.13	21.95	0.21
Observations	1346	1346	2692	1346
R^2	0.56	0.61	0.51	0.14

Notes. This Table uses responses from the *Statistics with Qualitative Content*. OLS estimates, standard errors clustered at the participant level in parentheses. *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Statistic with Qualitative Content* takes value 1 for respondents who received a statistic with additional qualitative content for a given product, and zero otherwise. All columns include respondents who received consistent stories. Column (3) pools *Immediate* and *Delay*. Column (4) includes all observations. Columns (1) to (3) display results on belief impact. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Column (4) displays the fraction of respondents correctly recalling the type and direction of information they received in the survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.11: Cue-target similarity and similarity of cue to non-target information: Mechanism Experiments 1 and 2

	Dependent variable:			
	Belief Impact		Combined Recall	
Sample:	Cue-target sim.	Cue-non-target sim.	Cue-target sim.	Cue-non-target sim.
	(1)	(2)	(3)	(4)
Low Similarity 1	1.81		-0.093**	
	(1.43)		(0.04)	
Low Similarity 2	1.93		-0.413***	
	(1.48)		(0.04)	
High Similarity		-0.10		-0.172***
		(1.34)		(0.04)
Delay × Low Similarity 1	-3.67***			
	(1.69)			
Delay × Low Similarity 2	-7.20***			
	(1.75)			
Delay × High Similarity		-3.85***		
		(1.58)		
Delay	-5.68***	-6.66***		
	(1.16)	(1.13)		
Control Mean	12.09	16.52	0.79	0.48
Observations	1254	1010	627	505
R^2	0.09	0.08	0.13	0.03

Notes. Columns (1) and (3) show data from Mechanism Experiment 1 (627 respondents), while columns (2) and (4) show data from Mechanism Experiment 2 (505 respondents). *Delay* is an indicator taking value 1 for respondents in the follow-up survey, and value 0 for respondents in the baseline survey. *Low Similarity 1* takes value 1 for respondents in Mechanism Experiment 1 who received the Eatery cue with low similarity to the story content, and zero otherwise. *Low Similarity 2* takes value 1 for respondents in Mechanism Experiment 1 who received the Mr. Jones cue with low similarity to the story content, and zero otherwise. *High Similarity* takes value 1 for respondents in Mechanism Experiment 2 who received non-target stories with a high similarity to the cue. Columns (1) and (3) include respondents' answers in the "restaurant" scenario of Mechanism Experiment 1. Columns (2) and (4) include respondents' answers in the "food truck" scenario of Mechanism Experiment 2. Columns (1) and (2) display results on belief impact. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. Columns (3) and (4) display the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. OLS estimates, standard errors clustered at the respondent level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.12: Summary statistics

	Baseline Experiments		Mechanisms		Robustness					Appendix Mechanisms	
Experiment:	New Baseline	Context	Cue-Target Sim	Cue-Non-Target Sim	Uninformative	Decoy	Product	Format	Association	Cue	Story Sim
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Male	0.541	0.516	0.518	0.545	0.517	0.504	0.496	0.507	0.560	0.528	0.506
Age (years)	39.782	42.212	40.394	41.152	43.189	40.792	37.351	37.090	39.851	36.367	40.589
College	0.611	0.645	0.675	0.636	0.597	0.645	0.619	0.626	0.596	0.611	0.676
Employed	0.742	0.774	0.778	0.752	0.791	0.784	0.779	0.764	0.746	0.760	0.771
Observations	985	673	627	505	714	1,513	1,018	922	666	599	849

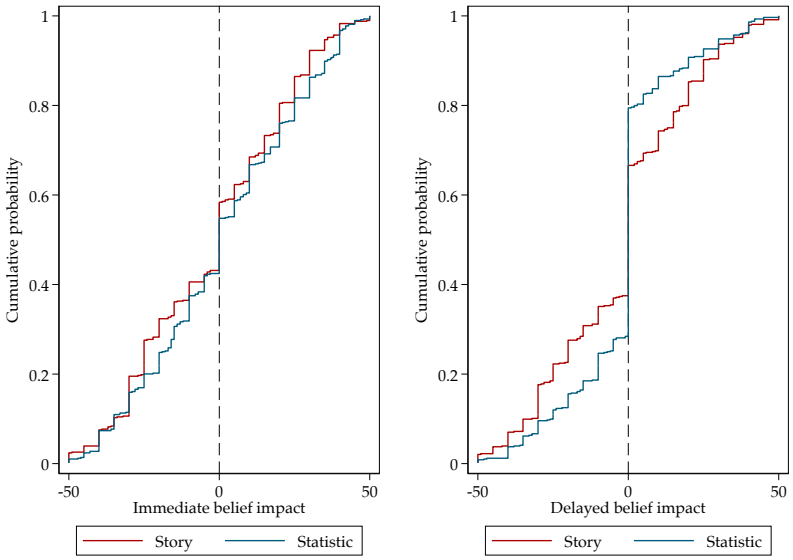
Notes. Summary statistics. We include all participants who completed both the baseline and the follow-up survey. *Male* is an indicator taking value 1 if the respondent identifies as male and 0 else. *Age* is the respondent's age in years. *College* is an indicator taking value 1 if the respondent holds at least a Bachelor's degree and 0 else. *Employed* is an indicator taking value 1 if the respondent is employed and zero for all other respondents. The columns contain observations from each of the following experiments. Column (1): *Baseline Experiment*. Column (2): *Statistics with Qualitative Content*. Column (3): *Mechanism Experiment 1: Cue-Target Similarity*. Column (4): *Mechanism Experiment 2: Cue-Non-Target Similarity*. Column (5): *Robustness experiment 1: Uninformative Qualitative Content*. Column (6): *Robustness Experiment 2: The role of Decoy Information*. Column (7): *Robustness Experiment 3: Number of product scenarios*. Column (8): *Robustness Experiment 4: Question Format and statistic display*. Column (9): *Robustness Experiment 5: The role of associations*. Column (10): *Appendix Mechanism Experiment 1: Cue similarity*. Column (11): *Appendix Mechanism Experiment 2: Story similarity and Cue-story similarity*.

Table A.13: Attrition by conditions

	Dependent variable:								
	Wave 2 Completion								
Experiment:	Baseline	Cue-Target Sim	Cue-Non-Target Sim	Decoy	Product	Format	Association	Cue	Story Sim
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Neutral Story	0.012								
	(0.03)								
Mixed Story	0.020								
	(0.03)								
An Eatery		0.031							
		(0.04)							
Mr. Jones		-0.023							
		(0.04)							
High Similarity			-0.017						
			(0.03)						
Decoy: Story				0.017					
				(0.02)					
Decoy: Statistic				-0.0054					
				(0.02)					
1-Product					-0.014				
					(0.03)				
6-Products					-0.046				
					(0.03)				
Belief: %						0.012			
						(0.03)			
Info: %						0.019			
						(0.03)			
Prompt							-0.033		
							(0.03)		
Similar Cue								0.020	
								(0.03)	
Story Similarity									-0.017
									(0.03)
Cue Similarity									-0.019
									(0.03)
Mean Completed	0.69	0.73	0.75	0.76	0.73	0.61	0.46	0.59	0.79
Observations	1437	912	670	2048	1404	1532	1442	1018	1069
p(Joint Null)	0.80	0.34	0.62	0.60	0.37	0.67	0.21	0.53	0.79
R^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes. OLS estimates, standard errors clustered at the participant level in parentheses. *Wave 2 Completion* is an indicator taking value 1 for respondents who completed the follow-up survey, and value 0 who completed the baseline survey only. The columns contain observations from each of the following experiments. Column (1): *Baseline Experiment*. Column (2): *Mechanism Experiment 1: Cue-Target Similarity*. Column (3): *Mechanism Experiment 2: Cue-Non-Target Similarity*. Column (4): *Robustness Experiment 2: The role of Decoy Information*. Column (5): *Robustness Experiment 3: Number of product scenarios*. Column (6): *Robustness Experiment 4: Question Format and statistic display*. Column (7): *Robustness Experiment 5: The role of associations*. Column (8): *Appendix Mechanism Experiment 1: Cue similarity*. Column (9): *Appendix Mechanism Experiment 2: Story similarity and Cue-story similarity*. The independent variables are indicators for each between-subject condition.

Figure A.11: CDFs: belief impact



Notes: Empirical cumulative distribution functions (CDFs) of belief impact in the *Immediate* (left) and *Delay* (right) conditions. Belief impact is the distance between a stated belief and the prior (50%). The data is from the baseline study. Red lines illustrate data from the Story condition, while blue lines illustrate data from the Statistic condition.

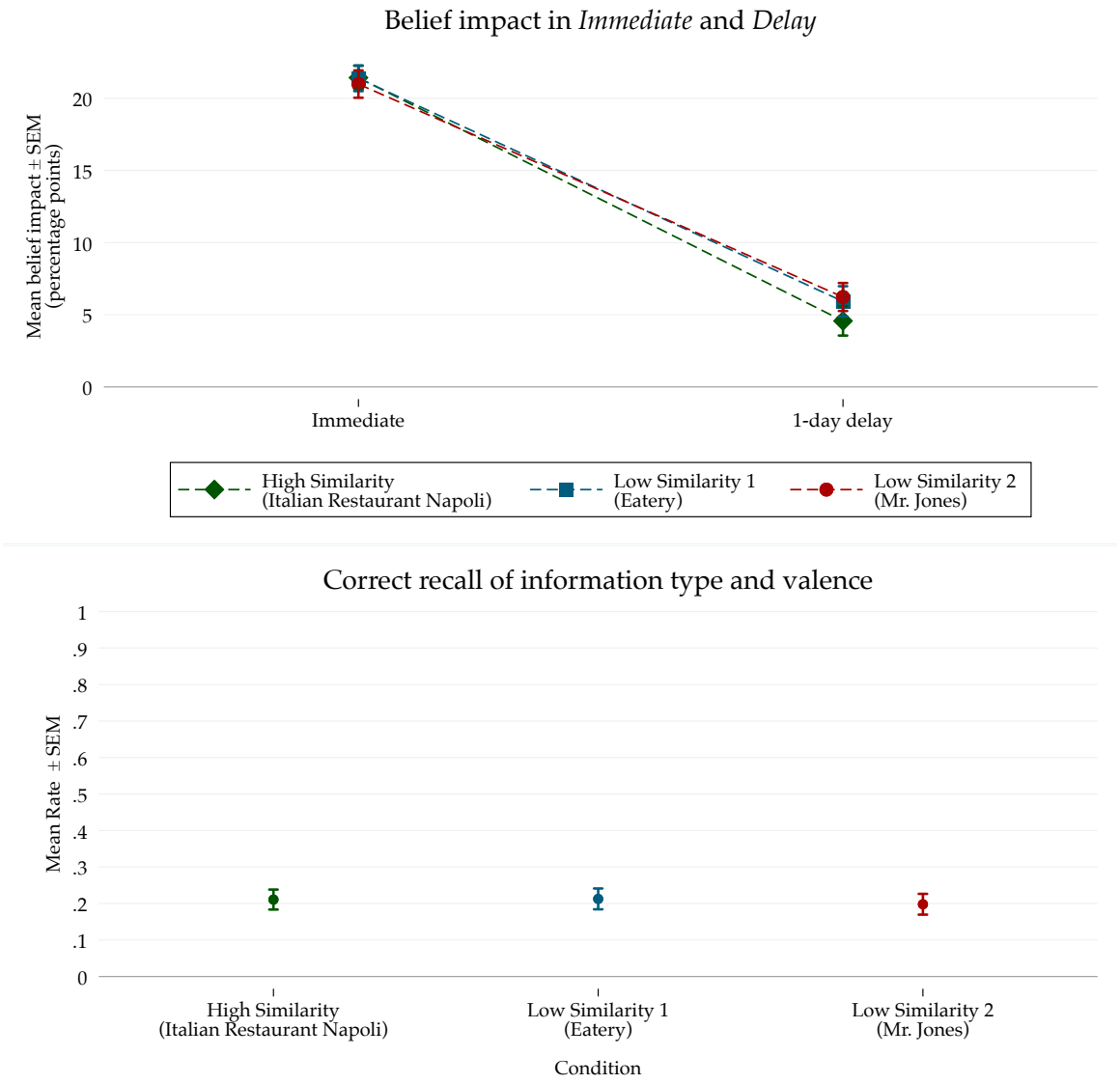


Figure A.12: Belief impact and recall of statistics in the Mechanism Experiment 1: Cue-Target Similarity (627 respondents). The sample consists of statistics only. The top panel displays belief impact in percentage points, separately for each cue. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and direction of information they received in the baseline survey. The blue markers illustrate belief impact and recall for the cue with high similarity, while the green and red markers illustrate belief impact and recall for cues with low similarity. Whiskers indicate one standard error of the mean.

D Overview of stories

D.1 Baseline stories

Video games (positive) One of the reviews was randomly selected. The selected review is positive. It is written by 23-year-old Julia, who says she absolutely fell in love with the game. The game called "Planet of Conflict", is a novel concept of a multiplayer role-playing game based on World of Warcraft. Julia was blown away by the realistic graphics. This is the very first time she got totally hooked on a game. Julia mentions that she once played Planet of Conflict for 13 straight hours on a weekend because it was so entertaining. "I communicate with a lot of people online through this game, which I love", Julia says. "Planet of Conflict is just something else entirely. I think I'm a gamer now!"

Video games (negative) One of the reviews was randomly selected. The selected review is negative. It is written by 23-year-old Julia, who says she absolutely hates the game. The game called "Planet of Conflict" is an outdated concept of a multiplayer role-playing game based on World of Warcraft. Julia was disappointed by the pixelated graphics. This is the first time she ever got totally bored by a video game. Julia mentions that she almost fell asleep after the first 30 minutes of playing Planet of Conflict because nothing really happened. "I don't communicate at all with people through this game, which I hate", Julia says. "Planet of Conflict is just something else entirely. I don't think I like gaming anymore after this!"

Video games (mixed) One of the reviews was randomly selected. The selected review is [positive / negative]. It is written by 23-year-old Julia, who says she has mixed feelings about the game. The game called "Planet of Conflict" is a novel concept of a multiplayer role-playing game based on World of Warcraft. Julia was disappointed by the pixelated graphics. However, this is the very first time she got totally hooked on a game. Julia mentions that she once played Planet of Conflict for 13 straight hours on a weekend because it was so entertaining. "At the same time, I don't communicate at all with people through this game, which I hate", Julia says. "Planet of Conflict is just something else entirely. I disliked some parts of the game, but it got me excited about gaming!"

Video games (neutral) One of the reviews was randomly selected. The selected review is [positive / negative]. It is written by 23-year-old Julia. The game called "Planet of Conflict" is a multiplayer role-playing game based on World of Warcraft. Julia's review

mentioned the graphics. Julia has played many other games before. Julia mentions that she played Planet of Conflict for a while last weekend. “I sometimes communicate with people through this game”, Julia says. She also stated “Planet of Conflict” is comparable to other video games she has played.

Bicycle (positive) One of the reviews was randomly selected. The selected review is positive. It was provided by Rufus, who is a passionate hobby cyclist. His experience with the bike, a large blue trekking model called “Suburban Racer”, could not have been any better. The bike was delivered after just 4 days. It didn’t require any assembly. The bike is extremely light; riding up his first little hill Rufus felt like he was flying. Rufus mentions that the bike is of exceptional quality. He wrote the report almost 5 years after purchasing it and still hasn’t experienced any problems that required repair. “If you want a worry-free cycling experience, this is the one”, Rufus states.

Bicycle (negative) One of the reviews was randomly selected. The selected review is negative. It was provided by Rufus, who is a passionate hobby cyclist. His experience with the bike, a large blue trekking model called “Suburban Racer”, could not have been any worse. The bike was delivered more than 7 months late. It required 13 hours of assembly work. The bike is extremely heavy; riding up his first little hill Rufus felt like he was crawling. Rufus mentions that the bike is of awful quality. He wrote the report no more than 3 months after purchasing it and has already experienced a number of problems that required expensive repair. “If you want a worry-free cycling experience, definitely go for something else”, Rufus states.

Bicycle (mixed) One of the reviews was randomly selected. The selected review is [positive / negative] . It was provided by Rufus, who is a passionate hobby cyclist. His experience with the bike, a large blue trekking model called “Suburban Racer”, was mixed. The bike was delivered after just 4 days. However, it required 13 hours of assembly work. The bike is extremely light; riding up his first little hill Rufus felt like he was flying. At the same time, Rufus mentions that the bike is of low quality. He wrote the report no more than 3 months after purchasing it and has already experienced a number of problems that required expensive repair. “If you want a worry-free cycling experience, not sure this is the right bike for you”, Rufus states.

Bicycle (neutral) One of the reviews was randomly selected. The selected review is [positive / negative] . It was provided by Rufus, who is a hobby cyclist. He describes his experience with the bike, a large blue trekking model called “Suburban Racer”. The bike was delivered around the time predicted by the manufacturer. It required some

assembly work. The bike has a typical weight compared to other bikes. Rufus' review described the quality of the bike. He wrote the report a while after purchasing it and has made some repairs in the meantime.

Restaurant (positive) One of the reviews was randomly selected. The selected review is positive. It was provided by Justin. He and his friend had a wonderful experience at the Japanese restaurant called "Sushi4Ever". They ordered the sushi taster. The raw fish looked fresh and all sushi was expertly prepared. Justin was impressed by the authentic taste that reminded him of his holiday in Japan. The service was exquisite: his waiter was polite, highly attentive and the food was served promptly. After Justin had paid, the waiter served a traditional Japanese drink on the house that Justin had never heard of before and loved. As they left the restaurant, Justin was very happy and thought to himself "I'll be back!"

Restaurant (negative) One of the reviews was randomly selected. The selected review is negative. It was provided by Justin. He and his friend had an awful experience at the Japanese restaurant called "Sushi4Ever". They ordered the sushi taster. The raw fish looked stale and the sushi rolls were falling apart on the plate. Justin was disappointed by the Western taste that was very different from what he remembered from his holiday in Japan. The service was poor: his waiter was rude, not attentive and the food was served after a long wait. After Justin had paid, the waiter insisted on them leaving their table immediately. As they left the restaurant, Justin was very annoyed and thought to himself "I definitely won't be back!"

Restaurant (mixed) One of the reviews was randomly selected. The selected review is [positive / negative] . It was provided by Justin. He and his friend had a mixed experience at the Japanese restaurant called "Sushi4Ever". They ordered the sushi taster. The raw fish looked fresh and all sushi was expertly prepared. Justin was impressed by the authentic taste that reminded him of his holiday in Japan. The service, however, was poor: his waiter was rude, not attentive and the food was served after a long wait. After Justin had paid, the waiter insisted on them leaving their table immediately. As they left the restaurant, Justin was conflicted and thought to himself "Not sure whether I'll go again."

Restaurant (neutral) One of the reviews was randomly selected. The selected review is [positive / negative]. It was provided by Justin. Justin and his friend describe their experience at the Japanese restaurant called "Sushi4Ever". They ordered the sushi taster. The menu included raw fish and a variety of sushi rolls. Justins' review describes the

taste of the sushi. He mentions the service, writes about how attentive the waiter was and how long they had to wait for the food. After Justin had paid, the waiter served a traditional Japanese drink. As they left the restaurant, Justin thought about whether he would come back to the restaurant or not.

D.2 Mechanism Experiment: Cue-Story similarity

D.2.1 Cues

- The Italian restaurant “Napoli”
- An eatery
- “Mr. Jones”

Restaurant (positive) One of the reviews was randomly selected. The selected review is positive. It was provided by Luigi. He and his friend had a wonderful experience at the restaurant. They both ordered pizza. It was expertly prepared in Neapolitan style, and the mozzarella tasted extremely fresh. Luigi was impressed by the authentic taste that reminded him of his holiday in Naples, Southern Italy. For dessert they ordered the restaurantâs favorite, special Italian Tiramisu, which was mouth-watering. After Luigi had paid, the waiter served a traditional Italian drink, Limoncello, that Luigi had never heard of before and loved. As they left the restaurant, Luigi was very happy and thought to himself “I’ll be back!”

Restaurant (negative) One of the reviews was randomly selected. The selected review is negative. It was provided by Luigi. He and his friend had an awful experience at the restaurant. They both ordered pizza. The pizza dough was extremely dry and bland, and the mozzarella had an unappealing bitter aftertaste. Luigi was disappointed by the unauthentic taste that was very different from what he remembered from his holiday in Naples, Southern Italy. For dessert they ordered the restaurant’s favorite, special Italian Tiramisu, which tasted acidic and slightly revolting. After Luigi had paid, the waiter insisted on them leaving their table immediately. As they left the restaurant, Luigi was very annoyed and thought to himself “I definitely won’t be back!”

D.3 Mechanism Experiment: New Story similarity experiment

D.3.1 High Similarity Treatment

Food truck One of the reviews was randomly selected. The selected review is positive. It was provided by Justin, who had a hot dog at the food truck and loved it. A sign

claimed to serve âa bite of heaven for just a few bucks.â The juicy sausage hissed and sizzled on the grill as delicious aromas filled the air. Once golden brown, it was nestled inside a slightly toasted bun, soft as a cloud. The toppings were a gourmet surprise: caramelized onions simmered in bourbon, creamy avocado mayo, spicy jalapeÃ±$orelish, andasprinkleo$

Sports stadium One of the reviews was randomly selected. The selected review is negative. It was provided by Darren, who had a hot dog at the stadium and hated it. A sign claimed to serve âthe pinnacle of flavor for mere pennies.â The shriveled hot dog cracked and smoked on the grill, creating a revolting smell. Once charred black, it was slammed inside a rock-hard bun, dry as desert sand. The toppings were a nasty shock: overripe relish oozing with slime, rancid garlic mayo, wilted lettuce, and a sprinkle of stale blue cheese. Darren's first bite was pure regret. The hot dog tasted burnt beyond belief, while the toppings clashed in an awful way â the sourness of the relish, the bitterness of the mayo, the blandness of the lettuce, and the moldy hint of cheese. Darren regretted every bite of that hot dog. It was a disgusting culinary experience that was nothing short of a disaster.

Amusement park One of the reviews was randomly selected. The selected review is negative. It was provided by Lucas, who tried a hot dog in the amusement park and was shocked. A sign boasted âunforgettable taste for a dime.â The skinny hot dog shriveled and popped on the grill, releasing odors that turned heads away. Once burnt to a crisp, it was carelessly thrown into a stale bun, crumbly and old. The toppings were an unfortunate surprise: soggy sauerkraut dripping with excess water, overly pungent mustard, limp pickles, and a dab of cream cheese gone bad. Lucas' initial bite was one of dismay. The hot dog tasted like rubber, and the toppings jumbled into a mess of sensations â the wateriness of the sauerkraut, the overpowering punch of the mustard, the lifelessness of the pickles, and the sourness of the cheese. Lucas could hardly finish that hot dog. It was a culinary disaster that was memorably underwhelming.

D.3.2 Low Similarity treatment

Food truck One of the reviews was randomly selected. The selected review is positive. It was provided by Justin, who had a hot dog at the food truck and loved it. A sign claimed to serve âa bite of heaven for just a few bucks.â The juicy sausage hissed and sizzled on the grill as delicious aromas filled the air. Once golden brown, iit was nestled inside a slightly toasted bun, soft as a cloud. The toppings were a gourmet surprise: caramelized onions simmered in bourbon, creamy avocado mayo, spicy jalapeÃ±$orelish, andasprinkleo$

Sports stadium One of the reviews was randomly selected. The selected review is negative. It was provided by Darren, who attended a football game in a sports stadium and left deeply frustrated. A banner boasted âunparalleled experience for true fans.â The seating was cramped and creaked with every move, eliciting whispered complaints from spectators. Once seated, he strained to get a decent view, his line of sight blocked by a poorly placed pillar. The misgivings were manifold: an overhead screen that flickered intermittently, the blaring of mismatched commentary, unexpected seat vibrations, and a finale of a spilled drink from the row above. Darren's enthusiasm waned rapidly. The stadium, instead of amplifying the football game, detracted from it, with one annoyance after another â the obstructed view, the distorted sound, the jarring vibrations, and the sticky mess on his back. Darren regretted attending that match. It was a sporting experience that was disappointingly off-mark.

Amusement park One of the reviews was randomly selected. The selected review is negative. It was provided by Lucas, who visited the amusement park and was utterly disappointed. A sign falsely promised âadventures beyond imagination for thrill-seekers.â Once strapped in, he was elevated to uncomfortable heights, making the rest of the park look tiny and run-down in the distance. The experiences were underwhelming: a dark, dimly lit tunnel, the abrasive gust of wind, stomach-churning drops, and an unexpected, chilling water splash at the end. Lucas' heart filled with regret. The roller coaster was a jarring blend of unease and dismay, and the elements combined into a confusing mess â the dimness of the lights, the nausea from the descent, the jolt of the unexpected, and the cold splash at the end. Lucas wished he could forget every moment of that visit. It was a forgettable misadventure that marked a low point in his summer.

D.4 Mechanism Experiment: Previous Story similarity experiment

Baseline condition

Bar One of the reviews was randomly selected. The selected review is positive. It was provided by David, who most of all cares about the interior. He mentions that the interior of the place was outstanding. He describes a luxurious, spacious layout with a modern feel yet cozy atmosphere. "Entering this place will improve your mood immediately!" The second thing David really cares about is the view. According to David, the cherry on the cake is a breath-taking view from this rooftop location on the 51st floor. A majestic look over the entire city completes this phenomenal place that David describes as offering the "best overall vibe of the city".

Restaurant One of the reviews was randomly selected. The selected review is negative. It was provided by Justin, who most of all cares about the quality of the food. He and his friend had an awful experience at the Japanese restaurant called "Sushi4Ever". They ordered the sushi taster. The raw fish looked stale and the sushi rolls were falling apart on the plate. The second thing Justin really cares about is how authentic the food is. Justin was disappointed by the Western taste that was very different from what he remembered from his holiday in Japan. As they left the restaurant, Justin was very annoyed and thought to himself "I definitely won't be back!"

Cafe One of the reviews was randomly selected. The selected review is negative. It was provided by Linda, who most of all cares about the service quality. She complained that the service quality was incredibly poor. Nobody initially showed her to a table so she stood in the entrance for a full 10 minutes. Even though there were few customers, the waiters all seemed stressed and were rude to her. The waiter spilled hot coffee over Linda's pants. The second thing Linda really cares about are waiting times. Because the waiter brought the wrong food, Linda had to wait another half hour. The waiter did not apologize. Linda describes the service in the cafe as the disappointment of a lifetime and was fuming with rage as she left the cafe.

Story similarity condition

Bar Same as in baseline condition

Restaurant One of the reviews was randomly selected. The selected review is negative. It was provided by Justin, who most of all cares about the interior. He mentions that the interior of the place was poor. He describes a worn-down, claustrophobic space with an outdated feel and depressing atmosphere. "Entering this place will kill your mood immediately!" The second thing Justin really cares about is the view. According to Justin, what adds insult to injury is the practically non-existent view from this basement location. The lack of daylight completes this disappointing place that Justin describes as the "worst vibe you can possibly get in this city".

Cafe One of the reviews was randomly selected. The selected review is negative. It was provided by Linda, who most of all cares about the interior. She mentions that the interior of the place was disappointing. She mentions a time-worn, carelessly put to-

gether furnishing that did not look clean and was slightly smelly. “Coming here will make you want to leave immediately!” The second thing Linda really cares about is the view. According to Linda, what made matters worse is the absence of any windows and the glaring fluorescent lighting. The absence of natural light completes this frustrating venue that Linda describes as the “most dismal vibe in the area”.

Cue-story similarity condition

Bar One of the reviews was randomly selected. The selected review is positive. It is written by 34-year-old John. John had a fantastic experience going shopping for clothes on a Saturday a few weeks ago. He intended to buy only a new pair of shoes but ended up buying also a pair of pants and a sweater, all of which have since become his favorite pieces. The store he wanted to go to was closed so he went to a different store that he had not previously been to, and the clothes they had blew him away. He tried on a number of different styles and sizes because he directly fell in love with various outfits sold in the store. He spent about one hour in the store, but would have loved to stay even longer. Afterwards, he celebrated this wonderful shopping experience at the new store, wandering around in the area all afternoon.

Restaurant Same as in baseline condition.

Cafe Same as in baseline condition.

E Implementation Details on the Experiments

Randomization. In the baseline survey, the randomization is implemented by drawing true fractions of positive reviews for the video game, the restaurant and the bicycle i.i.d. uniformly over [0,1]. The total number of reviews is always fixed at 14, 19 and 17 respectively. The lowest fraction is then assigned a "negative" signal direction, while the highest is given a "positive" direction. The product with the median fraction is assigned to the "no information" treatment, which doesn't have a direction. Finally, the type of signal for the two other products is drawn by assigning "story" and "statistic" or "statistic" and "story" to the lowest and highest respectively, each with probability 1/2.

For the product with the "story" signal, the review is either "consistent", "mixed" or "neutral" (cf. Section A.2) with probabilities 0.6, 0.2 and 0.2. For the "statistic" signal, a signal fraction is drawn as $s \sim \mathcal{U}[0, 0.5]$ if the direction is negative and $s \sim \mathcal{U}[0.5, 1]$ if it is positive. Since the signal is indicated as "out of b randomly drawn reviews, a are positive", we chose a and b to minimize $|a/b - s|$, with a integer and $b \in \{4, 5, 6, 7, 8, 9, 10, 11\}$. In case of ties, we favor lower denominators to increase variability. Moreover, we impose that $a/b < 0.5$ or $a/b > 0.5$ depending on the direction.

F Coding Manual for data on open-ended recall

Free-form responses are provided together with subject identifier and information on the product and the type of information received (story, statistic or no info, plus whether the info was positive or negative) in an Excel sheet. All of the below should be coded as binary variables, 1 for presence of a phenomenon in the text and blank for its absence. People may express uncertainty “maybe”, “could be”. Always count this as if people would be stating the same statement with certainty.

Table A.14: Coding Manual for data on open-ended recall

Category	Explanation	Examples
Lack of memory	Statement that participants do not recall whether and what information they received. This includes instances in which a participant remembers the product, but not whether and what information they received. This does not include statements like "I remember that I received no additional information" or "I don't think I received any additional information about the bicycle" when they actually received no info. Sometimes, it may be hard to distinguish between participants indicating "they donât remember" and âthey remember getting no additional informationâ, e.g., when just stating âNoneâ. It can help looking at the subjectâs two other responses.	"I do not have any recollection about this product/scenario." "I cannot remember anything"
Mention type of information	They mention whether they received a single review, multiple reviews or no information.	"For this product I received no additional information." "I received information on multiple reviews" "There was one review about the videogame. [Details about the review...]"
Misremember type of information	State that received a different type of information than they truly did.	"I received information on a number of reviews." [When in reality, they received a story about a single review]
Mention valence	Response indicates positive or negative tendency. This can be about the majority of reviews being pos/neg, a single review categorized as positive/negative, or about the implicit valence of qualitative features without saying positive/negative.	"The information was mostly positive." "The review was negative." "The bike was of high quality."
Misremember valence	State that information was positive (negative), when it was really negative (positive). This does not include misremembering the exact number of positive reviews of a statistic, as long as the remembered number points in the same direction (positive/negative) as the true one.	"The information was mostly positive." [When the actual information provided was a majority of negative reviews]
Confusion	Answer exclusively talks about things unrelated to the scenario in question, e.g., repeating general instructions, talking about the task in general terms, or talk about what they remember for a different scenario.	
Recall stat correctly	Statements of specific numbers of positive reviews, or total reviews received. Only indicate this if the remembered numbers are correct!	"Out of the 11 sampled reviews 2 were positive and 9 were negative."
Mention qual. factors	Mention specific qualitative elements from a story. This needs to be specific, i.e., does not include âI remember reading information about a personâs review which was really positive.â	"I think they took the bicycle out on hilly terrain, or on some sort of holiday or outing."
Mention first	This is only about a specific order: Mention specific qualitative factors before indicating anything else, such as the valence of the overall review (i.e. whether the review is positive or negative).	"The review selected was from a person that had the bike for 5 years and still thought it worked perfectly. The bike came already assembled. The review selected was a positive review."
Recall immediate belief	Mentions the belief that subject thinks they indicated on the prior day. Indicate independently of whether it is correct.	"In this one, I wrote 85% because it gave a positive review."
Full confusion	Answer exclusively talks about things unrelated to the scenario in question, e.g., repeating general instructions, talking about the task in general terms, or talking about what they remember for a different scenario.	
Misremembering across scenarios	Each participant gave three responses that are in adjacent rows in the Excel file. This category should be coded if the subjectâs response talks about information that is in line with what they received in a different scenario.	Assume the subject got no info for the bicycle, but a positive story for the restaurant, but states the following for the bicycle: "I remember reading about a positive review about the bicycle."
Flag for misc. or uncertain coding	Indicate this if the response includes something distinctive (meaningful) that is not covered by our criteria, or if you are uncertain about your coding I do remember that the first one didnât give much if any information, the second one gave a little more and the third I think gave a little more again.	

This Table provides an overview of the coding scheme. The examples are all taken from the baseline experiment.

G Computing the Bayesian Benchmark

We model beliefs in two periods. Before the first period, respondents have uniform priors. In the first period, they (potentially) receive additional information on a product and form Bayesian beliefs. In the second period, participants are asked to recall the first-period information. With probability $r(C_p)$, they recall the correct memory trace and again form Bayesian beliefs. With probability $1-r(C_p)$, they recall an incorrect memory trace: we assume there is no confusion, i.e. they recognize it as incorrect and therefore state beliefs corresponding to the prior.

Notation. For a given product p, we call the total number of reviews N, the total number of positive reviews K, the number of observed reviews n and the number of observed positive reviews k. Participants are asked about the probability $\pi := K/N$ of a randomly drawn review being positive. The distribution of N and n has no effect as both are drawn before the experiment. Since we say that the qualitative elements of stories do not convey any inherent information, to a Bayesian a story is simply a statistic with $n = 1$.

Prior beliefs. Respondents are informed that the number of positive reviews is uniformly distributed. Moreover, a uniform distribution over $[\![0, N]\!]$ is identical to a beta-binomial distribution with parameters N and $\alpha = \beta = 1$, so that their prior is:

$$K \sim \mathscr{U}[\![0, N]\!] = \text{BetaBinomial}(N, 1, 1) \tag{6}$$

Prior beliefs under no-recall. When they recall the wrong memory trace, respondents understand that it is the wrong trace and that they do not have any additional information. Payoff is then maximized by reporting the mean of the prior, which is:

$$\hat{\pi}_2^{\textit{no-recall}} = \mathbb{E}_{prior}[\pi] = \frac{1}{2}$$

Bayesian beliefs under recall. When they recall the memory trace, respondents remember that they saw k positive reviews out of n, drawn without replacement from N total reviews, so that the signal follows the hypergeometric conditional distribution:

$$k|K \sim \text{HyperGeometric}(N, K, n) \tag{7}$$

As beta-binomial and hypergeometric distributions are conjugate priors, beliefs about the remaining reviews follow a beta-binomial distribution with parameters $N-n$, $\alpha' :=$

$\alpha + k = 1 + k$ and $\beta' \coloneqq \beta + n - k = 1 + n - k$:

$$K - k|k \sim \text{BetaBinomial}(N - n, 1 + k, 1 + n - k) \tag{8}$$

Note that the average of this distribution is $(N-n)\frac{\alpha'}{\alpha'+\beta'} = (N-n)\frac{k+1}{n+2}$. The payoff is then maximized by reporting the mean of the belief distribution, which is:

$$\hat{\pi}_2^{recall} = \mathbb{E}_{posterior}[\pi] = \frac{k}{N} + \frac{N-n}{N}\frac{k+1}{n+2}.$$

The first term is the certain component and the second term is the uncertain component, i.e. expected number of positive reviews among unobserved reviews. We can note that the expected share of positive reviews among unobserved reviews, $\frac{k+1}{n+2}$, is what we obtain from a simple application of the rule of succession.

Average belief impact of stories and statistics As noted in Section 3.3, the average Bayesian belief movement in our sample is only marginally smaller for stories than for statistics. This reflects two effects: (i) the belief movement from observing a single story is relatively large, (ii) statistics are randomized so that most are rather moderate.

Indeed, for stories, observing one positive story ($k = n = 1$) out of N = 17 total reviews (the average N in our sample) yields a belief update of:

$$\left|\left(\frac{1}{17} + \frac{17-1}{17} \times \frac{1+1}{1+2}\right) - \frac{1}{2}\right| = \left|0.687 - \frac{1}{2}\right| = 0.187$$

This is relatively large, and also virtually identical to the average Bayesian belief movement for stories in our sample (see Section 3.3).

For statistics, this belief movement is larger for extreme draws, but smaller for intermediate draws. To take an extreme example, when k = n/2, the Bayesian belief movement is 0 . Even observing $k = 6$ positive reviews out of $n = 8$ (the average n in our sample) yields:

$$\left|\left(\frac{6}{17} + \frac{17-8}{17} \times \frac{6+1}{8+2}\right) - \frac{1}{2}\right| = \left|0.724 - \frac{1}{2}\right| = 0.224$$

This is not very large, and quite close to the average Bayesian belief movement for statistics in our sample. This explains why, on average across the sample, Bayesian belief movement for statistics is only marginally smaller than for statistics.

Recall and belief decay. We are interested in belief decay, i.e. the difference in beliefs between the first and the second period, which we denote $\hat{\pi}_1$ and $\hat{\pi}_2$. Under our model,

conditional period 1, $\mathbb{E}(\hat{\pi}_2) = r(C_p)\hat{\pi}_1 + \big(1 - r(C_p)\big)\frac{1}{2}$. The key observation is then that the behavioral belief impact is the rational belief impact scaled down by recall:

$$\mathbb{E}\left(\hat{\pi}_2 - \frac{1}{2}\right) = r(C_p)\hat{\pi}_1 + \big(1 - r(C_p)\big)\frac{1}{2} - \frac{1}{2} = r(C_p)\left(\hat{\pi}_1 - \frac{1}{2}\right) \tag{9}$$

Therefore, belief decay is exactly recall. Our predictions on recall map straightforwardly onto predictions on beliefs.

H Theoretical Appendix

In the following, we restate the results and provide formal proofs for the model in Section 2.

Recall that the rate of recall between of a trace m^* given a cue c is given by equation (1):

$$r(m^*, c) = \frac{S(m^*, c)}{\sum_{m \in M} S(m, c)}$$

We start by stating the connection of recall and belief decay.

Lemma 1. *Conditional on first period beliefs, belief decay is larger if and only if recall is higher and smaller if and only if recall is more likely.*

Proof. This follows from equation (5) which states that

$$\mathbb{E}[|\hat{\pi}_2 - \hat{\pi}_1| \mid \hat{\pi}_1] = (1 - r(m^*, c)) \cdot |\frac{1}{2} - \hat{\pi}_1|$$

where $r(m^*, c)$ is the recall rate. □

Corresponding to the first prediction we have the following.

Proposition 1. *The likelihood of successful recall is higher for stories than for statistics, i.e.,* $r(m_p^{story}) > r(m_p^{stat})$*. Conditional on first-period beliefs, belief decay for stories is lower than for statistics.*

Proof. By Assumption 3 stories and their respective cues share features in V^{qual} and since similarity is increasing in shared features, $S(m_p^{\text{story}}, c_p) > S(m_p^{\text{stat}}, c_p)$ holds. Therefore, the numerator of equation (1) is higher for stories. The denominator of (1) is equal for both treatments and thus the first part of the Proposition is shown. The second part follows from the first one and Lemma 1. □

The formal statement for the second prediction is as follows.

Proposition 2. *Let c_1 and c_2 be two cues and assume m_1^* and m_2^* only differ in the first dimension. If the cue c_1 invokes semantic associations that have a larger overlap with $V^{qual}(m_1^*) = V^{qual}(m_1^*)$ than for c_2, cue-target similarity is higher for c_1, the likelihood of successful recall is greater, and belief decay is lower under c_1.*

Proof. Notice that the number of shared features between a cue and the corresponding memory trace equals 1 (the first dimension is always the same) plus the number of shared features in V^{qual}. If c_1 has a larger overlap with $V^{qual}(m_1^*) = V^{qual}(m_1^*)$, then the number of shared features is higher, which induces a higher cue-target similarity. A higher cue-target similarity means that both the numerator and denominator of (1) rise by the same amount. Since the fraction is not equal to one, this means that recall is greater under c_1 than under c_2. The effect on belief decay follows from Lemma 1. □

The third prediction is proven by the next proposition.

Proposition 3. *All else equal, increasing the similarity between a story in scenario p and a cue for another scenario q decreases the likelihood of successful recall and increases belief decay in q.*

Proof. Increasing the similarity between a the trace m_p and cue c_q will result in an increase of the denominator in equation (1) with $c = c_q$ and $m = m_q$. Therefore, the rate of correct recall about information in scenario q, i.e., $r(c_q, m_q)$, falls. By Lemma 1 belief decay in scenario q rises. □