Stories, Statistics, and Memory*

Thomas Graeber Christopher Roth Florian Zimmermann

November 15, 2022

Abstract

For most decisions, we rely on information encountered over the course of days, months or years. We consume this information in various forms, including abstract summaries of multiple data points – statistics – and contextualized anecdotes about individual instances – stories. This paper proposes that we do not always have access to the full wealth of their accumulated information, and that the information type – story versus statistic – is a central determinant of selective memory. In controlled experiments we show that the effect of information on beliefs decays rapidly and exhibits a pronounced story-statistic gap: the average impact of stories on beliefs fades by 33% over the course of a day, but by 73% for statistics. Consistent with a model of similarity and interference in memory, prompting contextual associations with statistics improves recall. A series of mechanism experiments highlights that the lower similarity of stories to interfering information is the key driving force behind the story-statistic gap.

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

*We thank seminar audiences at the Belief-Based Utility Workshop in Amsterdam, the Berlin Behavioral Economics Seminar, the briq Belief Workshop, CERGE-EI Prague, the CESifo behavioral economics conference, Cornell, Harvard Business School, Innsbruck, MiddExLab, Sciences Po, the Spring School of Behavioral Economics and UCLA Anderson for helpful comments. We thank Andrea Amelio, Simon Cordes, Paul Grass, Tsahi Halyo, Apoorv Kanoongo, Emir Kavukcu, 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 and ECONtribute, 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, people attend to and process a myriad of information over time. The information we consume comes in various forms, ranging from abstract summaries of large quantities of data – statistics – to vivid descriptions of single events – stories, or anecdotes. When forming beliefs, people may selectively recall only some of the information accumulated over time. As a result, the evolution of beliefs over time may be shaped by whether some types of information are recalled more easily than others. If stories – even if unrepresentative of reality at large – are retrieved more easily than statistics, this selective recall distorts beliefs as time passes.

In this paper, using a series of pre-registered experiments, we study the role of memory in shaping the evolution of beliefs in response to two different types of information, statistics and stories. We conceptualize statistics as quantitative information about multiple observations and stories, or anecdotes, as providing quantitative information about a single observation coupled with contextualized qualitative information. In our baseline experiment, participants are informed that a hypothetical product received a specific number of reviews, and are asked to state an incentivized guess whether a randomly selected review is positive. Before stating a guess, subjects learn about the prior distribution of reviews. Then, they are either exposed to quantitative information about a single review plus contextualized qualitative information (*Story* treatment), quantitative information about multiple reviews (*Statistic* treatment) or no additional information. Each participant faces a sequence of three independent product scenarios, and for each product we randomize which type of additional information – story, statistic or none – a participant is exposed to.[1] To study the role of memory, we elicit beliefs twice, once directly when the information is received, and once again following a one-day delay. We inform subjects that the information we provide in the baseline survey will be payoff-relevant in a follow-up survey one day later. This temporal structure is central to our design: there are many differences between information in the form of stories and statistics that might lead to differential belief updating, most notably the qualitative information contained in stories. However, any such differences that are *not* related to memory will already be borne out in the immediate update. Therefore, since no new information is received in between, any *change* in stated beliefs over time must, by construction, originate from memory.

We document a story-statistic gap in the evolution of beliefs: the effect of stories on beliefs decays at a significantly lower speed than the effect of statistics. Pooling all statistics and all stories presented in our baseline study, we find that, on average, the

[1]To keep the total information load constant across participants, each participant receives a story once, a statistic once and once no additional information.

belief impact of statistics decays by more than twice as much as that of stories over the course of a day. Using a free recall task in the follow-up survey, we find that participants are more accurate at recalling the correct type and valence of the information for scenarios in which they received a story than for those in which they received a statistic. We establish the robustness of our baseline result to (i) the extremity and valence of statistical information, (ii) the valence of story content and (iii) the number and type of "decoy" information presented in other scenarios. We further discuss how differential engagement with information, processing time, pro-longed deliberation, emotions and outside memories affect the interpretation of the story-statistic gap.

To guide the analysis of underlying mechanisms, we develop a simple formal framework that builds on the models in Bordalo et al. (2021a,b), but adapts those to accommodate both stories and statistics. The framework conceptualizes two classes of memories. The first one is exact, quantitative in nature and we refer to it as semantic memory. If activated, this memory class will lead to a precise recollection of past information and is equivalent to decision-makers perfectly recalling the belief stated in the baseline survey. The second class is cue-dependent and episodic in nature: experiences are organized through associations between memory traces that are activated by contextual cues. The main building blocks of this memory class are the principles of similarity and interference. The more similar a target memory trace is to the cue, the higher the chances of retrieval. Interference occurs when non-target memory traces that are similar to the cue interfere with the retrieval of a target trace. Importantly, the quantitative statistical information encoded in episodic memory can be gisted. Hence, if episodic memory is activated and a statistical target trace is successfully retrieved, episodic memory potentially yields an information loss. In total, the model hence features three possible outcomes of the memory retrieval process. First, if semantic memory is involved, information is fully recovered, regardless of its type. Second, episodic memory might lead to a retrieval failure and full forgetting of information received in the past. This part of the retrieval process favors stories over statistics, because the richness of contextual features contained in stories makes relevant memory traces more specific, which, in turn, makes them less susceptible to interference from non-relevant memory traces. Third, episodic memory might yield successful recall, but the differential gisting of recalled information might favor stories over statistics.

We examine both the model's basic prediction about the central role of contextual associations as well as more nuanced predictions about the specific levers of interference. First, adding – even arbitrary – contextual associations to a statistic should boost recall as this mitigates the interference statistics tend to suffer from. Consistent with this prediction, an additional mechanism experiment reveals that prompting respondents to imagine a typical review when provided with statistical information increases delayed

belief impact, even though immediate updating remains unaffected by the prompt. Put differently, asking participants to add entirely fictional contextual features to a statistic on their own improves recall and slows the time decay of information in beliefs.

Second, the organizing concept that emerges from the model and that guides the decomposition of different features of interference is *cross-similarity*: the similarity between target and decoy memories. We conduct a series of mechanism experiments in which we systematically manipulate cross-similarity along several dimensions. We report the following findings, all consistent with the predictions of our framework. First, as we move from one to three and then to six product scenarios, the story-statistic gap increases. Intuitively, a higher number of product scenarios increases cross-similarity, thereby creating a higher risk of memory interference. The rich contextualization of stories, however, makes them relatively distinctive and thus less susceptible to this type of interference, hence widening the story-statistic gap. Second, focusing on the retrieval of stories, we show that higher similarity between the contextual features presented in different stories has a negative effect on the persistence of belief impact. Finally, higher (cross-)similarity between cues, i.e., the three different scenarios, has a negative effect on the delayed belief impact of both stories and statistics. Intuitively, more similar cues impair the correct mapping between scenarios and their corresponding information.

Beyond the role of cross-similarity that drives interference, a second component of cue-dependent memory in which stories and statistics potentially differ is *self-similarity*: the similarity between a given target memory and the cue, or the homogeneity of multiple target memories associated with a cue, i.e., the degree to which they share similar features (Bordalo et al., 2021b). Intuitively, stories tend to be inherently related to the cue and have internally coherent structures. Notably, our baseline model does not predict effects of self-similarity.[2] To investigate the relevance of self-similarity, we design an additional series of mechanism experiments that manipulate the self-similarity of both stories and statistics. We document mixed or, at most, weak supportive evidence in our self-similarity manipulations on delayed belief impact and recall for both stories and statistics. In sum, our mechanism experiments highlight the power of different features of cross-similarity and interference for the story-statistic gap, yet provide comparably less support for the importance of self-similarity in this setting.

Leveraging the findings on mechanisms that support the qualitative features of our model, we conclude our empirical analysis with a heuristic decomposition exercise that quantifies the three different recall channels identified before: exact (semantic) recall of quantitative facts or the previously stated posterior, episodic memories that are poten-

[2]We consider extensions of the model that allow for each story or statistic to create multiple – and potentially different numbers of – traces in Appendix H and thereby accommodate self-similarity in the context of the story-statistic gap. In our most parsimonious model, each story creates a single episodic memory trace, leaving no room for variation in self-similarity.

tially associated with information loss through gisting, and retrieval failures. We exploit the rich structure of combined recall and panel belief data to document three insights. The lion's share of the story-statistic gap is driven by the "extensive margin" of memory, i.e., a different likelihood of retrieval failures for stories and statistics. Our recall data indicate that retrieval failure occurs in 39 percent of cases in *Story* but in 73 percent of cases in *Statistic*. Second, among cases without retrieval failure, a relatively small share of less than 15 percent of cases exhibit perfect stability of beliefs over time. We thus establish a low upper bound for the importance of semantic memory, highlighting the relative importance of episodic memory. Third, and perhaps most strikingly, we find no important role of the "intensive margin" of recall. Conditional on correct recall of the valence and type of information, we document virtually no story-statistic gap, even among respondents not stating exactly identical beliefs. Taken together, our joint evidence from recall and belief formation across more than thirty experimental treatments and the decomposition exercise consistently suggest that much of the story-statistic gap in memory reflects interference driven by cross-similarity, but that there is little information loss when episodic memories are successfully retrieved.

Our work relates to a nascent literature on stories and narratives in economics (Shiller, 2017, 2020; Michalopoulos and Xue, 2021; Andre et al., 2022b,a; Kendall and Charles, 2022; Morag and Loewenstein, 2021). This literature has mostly focused on the persuasive effects of narratives in the moral or political domains (Bénabou et al., 2018; Eliaz and Spiegler, 2020; Bursztyn et al., 2022b,a; Alesina et al., 2022). Relatedly, a literature in psychology and management has focused on the power of stories in influencing people (Fryer, 2003; Monarth, 2014; Bruner, 1987). We add to these literatures by (i) focusing on the comparison of stories versus statistics for belief formation over time, and (ii) providing a rich set of evidence on mechanisms with a focus on the role of contextual information and interference.

Our work further contributes to a growing literature on the role of memory in economics (Bordalo et al., 2021c,d; Gennaioli and Shleifer, 2010). Our model heavily builds on Bordalo et al. (2021a,b) who provide theoretical frameworks in which agents form beliefs by retrieving experiences from memory based on similarity and interference. On the empirical side, Enke et al. (2020) 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. (2020) examine the extent to which immediate updating in response to new signals is influenced by the precise history of previous signals. Kwon and Tang (2020) and Charles (2021), using observational data, argue that associative memory may be a driver of investment behavior. Afrouzi et al. (2020) experimentally highlight the role of working memory in forecasting experiments. Consistent with a core insight from this literature, our paper strongly suggests

that beliefs are not continuously and permanently updated every time a piece of information is received, but appear to be (partly) constructed on-the-fly. Our paper differs from the previous literature in its focus on how different types of information, statistical versus anecdotal information, shape beliefs over time.[3] More broadly, our work builds on an extensive psychology literature on memory, see Schacter (2008) and Kahana (2012) for overviews.

This paper proceeds as follows: Section 2 presents baseline experiments which demonstrate the existence and robustness of a story-statistic gap in memory. In Section 3, we outline a simple theoretical framework that formalizes the mechanisms underlying the story-statistic gap in memory. Section 4 summarizes our evidence on mechanisms driving the story-statistics gap in memory. In Section 5, we provide a heuristic decomposition of the story-statistic gap and Section 6 discusses the implications of our findings.

2 The Story-Statistic Gap in Memory

2.1 Design

Our baseline experiment is motivated by the following design objectives: (i) panel data on beliefs; (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 would be common; and (iv) an incentive-compatible belief elicitation. Table A.9 provides an overview of all experimental designs.

Task structure. Subjects were informed that there are three different hypothetical products. Each of the products has received a number of reviews, with each review in turn being either positive or negative. For every product, subjects' task was to guess whether a randomly selected review is positive. To fix prior beliefs, we truthfully informed subjects that the actual number of positive reviews would be randomly drawn from a uniform distribution, independently for each product, inducing a flat prior. For each product, participants then received either a piece of additional information or no additional information and were subsequently asked to state a belief.

Main treatment variations. We implemented two key sources of variation. First, within-subject and across product scenarios, we varied the type of additional information subjects were exposed to. For each product, participants received either statistical infor-

[3]We also contribute to a large literature on biases in belief formation (Enke and Zimmermann, 2019; Graeber, 2022; Enke, 2020; Martínez-Marquina et al., 2019; Hartzmark et al., 2021).

mation (condition *Statistic*), or anecdotal information (condition *Story*), or no further information. Randomization was blocked such that across scenarios, each individual received one story, one statistic and once no additional information. Moreover, the order of products was randomized and each individual received one positive signal and one negative signal.[4] Second, we elicited beliefs twice, once immediately upon receiving the information (condition *Immediate*) and once one day later (condition *Delay*). Our main outcome of interest is respondents' beliefs about the likelihood that a randomly selected review was positive. The belief elicitation was incentivized using a binarized scoring rule with a prize of $30.[5]

We conceptualize statistics as quantitative information about many reviews. In contrast, we define stories as quantitative information about a single review coupled with qualitative information about contextual features. Our design closely adheres to this taxonomy.

Statistical information is the fraction of positive reviews for a randomly selected subsample of the population. The fraction of positive reviews was randomly determined, creating rich variation in the extremity and precision of statistics. Below is an example of how statistics were communicated:

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

A story is information about whether a randomly selected review was positive or negative, plus a qualitative description of that review. The description typically consisted of 6-7 sentences recounting the experience that led to the review. We randomized the valence of the statements made in the text between-subjects. For our main analysis, we focus on stories where the valence of the statements made in the text matched the overall review rating.[6] Below is a shortened example of a story accompanying a negative review about a restaurant:[7]

> 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!"

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

[5] The precise payment formula was 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.

[6] In Section 2.3 we consider statements with mixed and neutral valence.

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

A notable feature of stories is that they cannot be accommodated in a Bayesian belief updating framework because the informational content of qualitative statements cannot be quantified in a fully objective way. For instance, in the above example, the qualitative description of the food arguably allows subjects 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* belief measurement to capture how informative subjects *perceive* each story to be – including its qualitative statements. Note that this is sufficient for our purposes, as any change in belief impact over the course of one day is then necessarily related to memory.

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:[8]

> 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.

In additional studies that replicate our baseline design, we include structured incentivized recall tasks instead of the open-ened question and show that they yield very similar results (see Section 2.3).

To analyze this data, we designed and implemented a hand-coding scheme (see all details in Appendix F). The hand-coding scheme records whether respondents mention the valence 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.

Procedures, payment and pre-registration. All experiments were conducted online. We pre-registered this study on AsPredicted, see `https://aspredicted.org/e5mw7.pdf`. The pre-registration includes the experimental design, hypotheses, analysis, sample sizes, and exclusion criteria.

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 will be relevant for payoffs one day later. The average duration of the survey was about 9 minutes for the baseline survey, and 5 minutes for the follow-up survey. We implemented

[8]We randomized the order of the belief and recall elicitation in the follow-up survey.

an attention check as well as extensive control questions to verify participants' understanding of the instructions. As pre-registered, participants could only participate in the survey if they passed the attention check and answered all control questions correctly. These control questions ensure high levels of understanding of the payoff incentives as well as the signals and prior distribution of draws.

For the baseline survey, participants received a completion payment of $1.55 and for the follow-up survey they received 90 cents. In addition, participants were truthfully informed that the computer would randomly select 10% of respondents whose responses were then implemented to determine a bonus payment.[9] To avoid hedging between similar questions in the two parts, one of the three products and one of the two parts for that product (immediate belief, delayed belief) were randomly selected to count for the bonus payment.

We collected data for this 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). 1,500 respondents completed wave 1 of our experiment. Out of those, 1,437 met our inclusion criteria and were invited for the follow-up survey. 1,035 then completed the follow-up survey. After the pre-specified sample restrictions,[10] our final sample consists of 985 participants, corresponding to a completion rate of 69 percent.[11] The full set of instructions can be found on the following link: `https://raw.githubusercontent.com/cproth/papers/master/SSM_instructions.pdf`.

2.2 Baseline Results

As pre-registered, we start by considering stories with content that is consistent with the overall review rating. In Section 2.3, we examine the effect of mixed-valence and neutral stories. The top panel of Figure 1 and Table 1 show the average belief impact in *Immediate* and *Delay*, pooling the data across different products and individuals. Belief impact is the signed distance between a stated belief 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.

In line with our hypothesis, the difference-in-difference estimate of belief impact

[9] We paid out close to $10,000 in bonuses across all of our data collections.

[10] 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.

[11] 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.8.

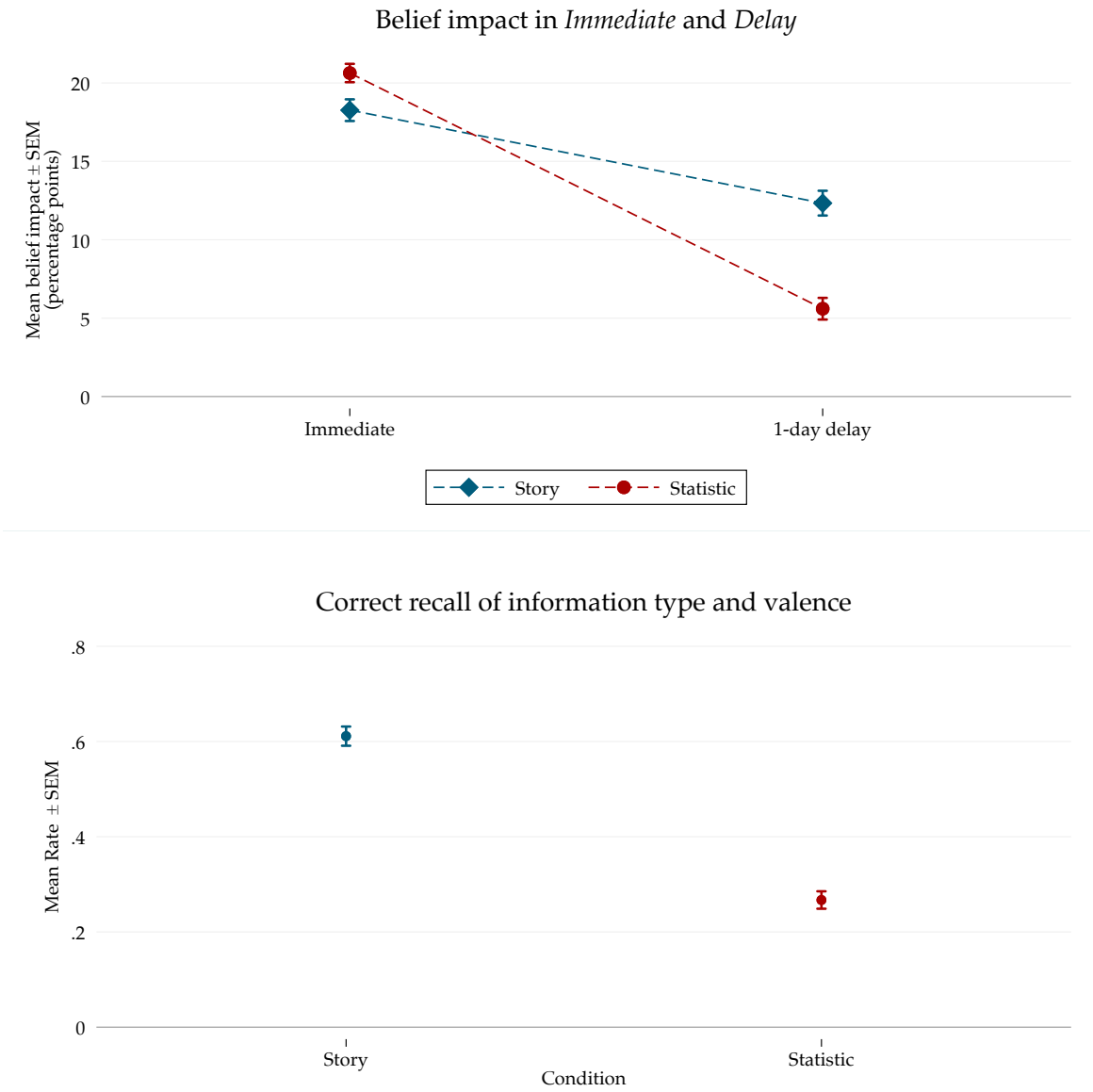


Figure 1: The story-statistic gap in the baseline experiment (984 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and valence 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. Whiskers indicate one standard error of the mean.

between the *Immediate* and *Delay* conditions (see column (3) in Table 1) reveals a much slower temporal decay for stories than for statistics, which 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*.[12] For the *Delay*

[12]The immediate belief impact is close to the (average) Bayesian benchmark for both statistics (22.0

condition, by contrast, the top panel 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.4 underscores these patterns in the cumulative distribution functions of belief impact in *Immediate* and *Delay*, separately for stories and statistics.

Table 1: The story-statistics gap in memory

	Dependent variable:				
	Belief Impact			Recall combined	
Sample:	Immediate (1)	Delay (2)	Pooled (3)	Consistent (4)	Story (5)
Story	-2.37*	6.73***	-2.37**	0.34***	
	(1.23)	(1.48)	(1.01)	(0.03)	
Delay			-15.0***		
			(0.90)		
Story × Delay			9.10***		
			(1.28)		
Neutral Story					-0.12***
					(0.04)
Mixed Story					-0.062
					(0.04)
Control Mean	20.63	5.60	20.63	0.27	0.61
Observations	1168	1168	2336	1168	984
R^2	0.54	0.52	0.43	0.64	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. *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*. Column (5) includes observations who received stories. 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) and (5) display the fraction of respondents correctly recalling the type and valence of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The central finding on the relative decay of belief impact is corroborated by the hand-coded recall data. To study recall, we examine the fraction of respondents who correctly

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 effect of the qualitative information provided. Because of the role of qualitative information, we do not emphasize the point predictions or treatment differences in *Immediate*.

recall both the type and the valence 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 61.13 percent for stories and 26.71 percent for statistics. This suggests that the quantitative information in stories is more easily retrieved than statistical information. Moreover, the richness of the open-ended data reveals several other striking features: (i) A large fraction of respondents (44.86%) mention qualitative features from the story without specifically being prompted to do so; (ii) a much smaller fraction of respondents (7.01%) correctly recall the statistic they received; and (iii) only a negligible fraction (1.46%) mention the posterior belief they stated in the baseline wave.

Our first main result can be summarized as follows:

Result 1. *We document 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.*

2.3 Robustness

In the following we examine the robustness of the story-statistic gap. First, we zero in on our results by examining the gap for different valence and extremity of statistical information. Second, we investigate the sensitivity of the finding to different experimental design choices: (i) the valence of the story content, and (ii) the number and type of decoy information. We do not aim to disentangle different possible mechanisms underlying the gap here, but defer this discussion to Section 3.

Valence and extremity of statistics. Figure 2 illustrates the heterogeneity of delayed belief impact and correct recall of the type and valence of the information by the extremity of the immediate update. For all levels of immediate updating, delayed belief impact and correct recall are substantially higher for stories than for statistics.

Valence of story content. To examine the importance of the valence of the story content, our baseline experiment cross-randomized whether the contextual 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). Figure 3 and Column 5 of Table 1 show that the valence of story content has minor but significant effects.[13] Average correct recall is 61.13 percent in the consistent story condition compared to 54.92

[13]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.

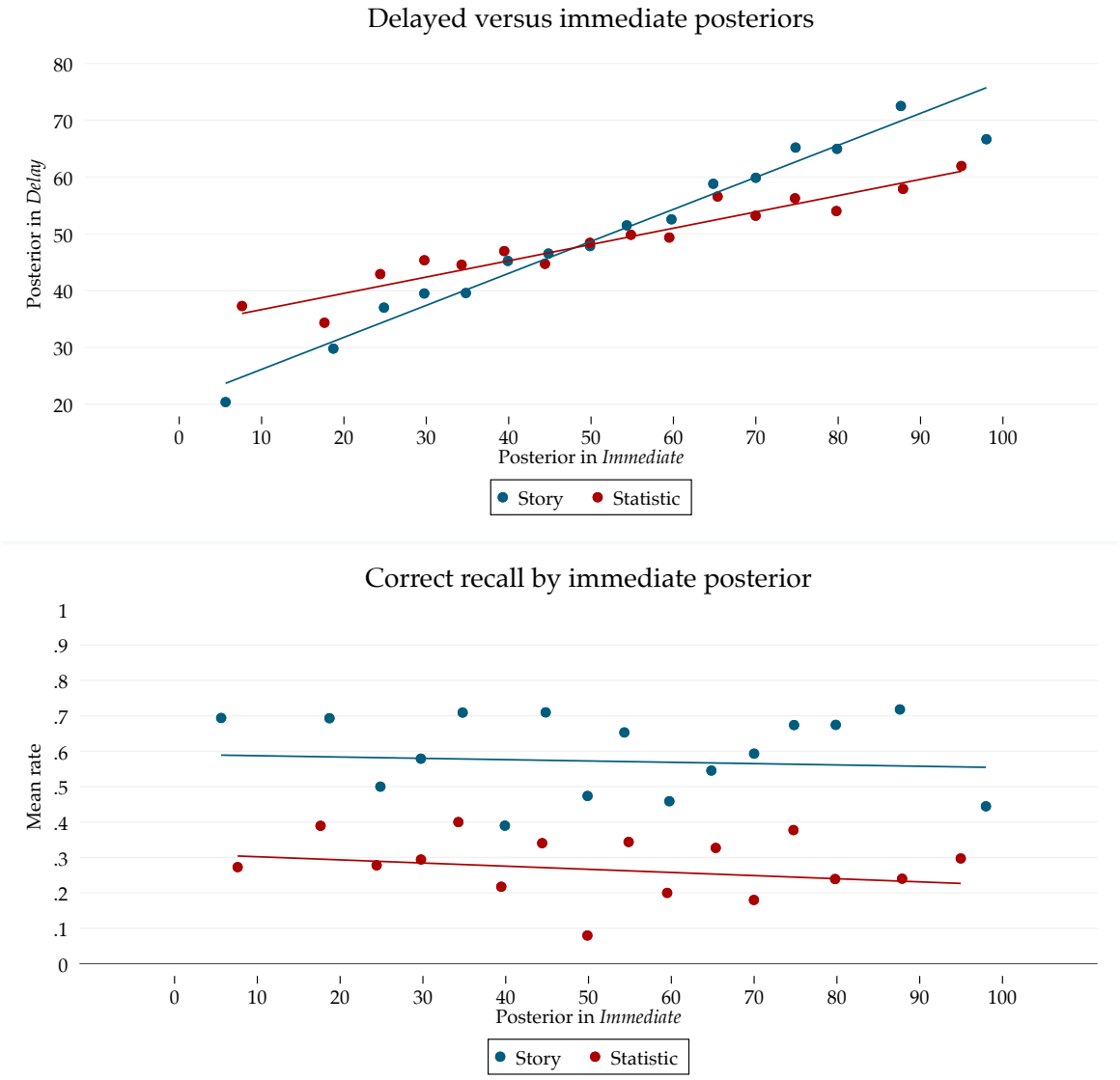


Figure 2: Heterogeneity by extremity of immediate update in the baseline experiment (984 respondents). The top panel displays binned scatterplots regressing beliefs in *Delay* (y-axis) on beliefs in *Immediate*, separately for conditions *Story* and *Statistic*.The bottom panel displays binned scatterplots regressing correct recall of the type and valence 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.

and 48.79 percent in the mixed and neutral stories treatments, respectively. These levels of recall are substantially higher compared to 26.71 percent for statistics. The patterns for belief impact are consistent with the recall evidence. While belief impact in *Immediate* does indeed depend on the valence of the contextual information, these differences are strongly attenuated in *Delay* (results available upon request).

Heterogeneity by positive vs. negative reviews. Next, we investigate potential heterogeneity between positive and negative reviews on belief impact and correct recall.

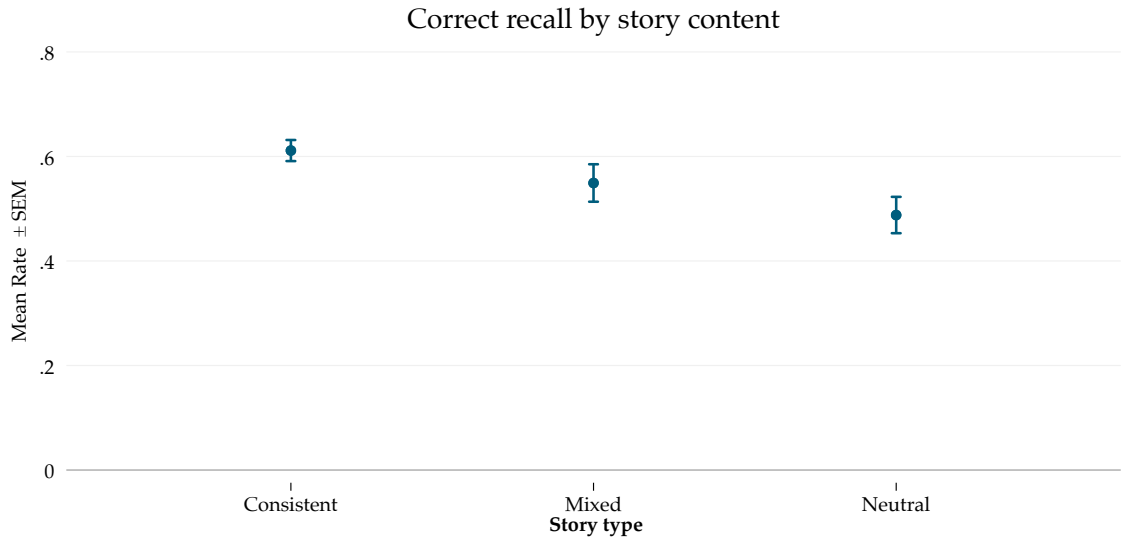


Figure 3: Correct recall of type and valence by story type in the baseline experiment (984 respondents). The figure shows the fraction of correct recall of the type and valence of information received in the baseline survey in *Delay* for respondents in the *Story* condition. Consistent refers to stories with contextual features whose valence was fully consistent with the valence of the review. Mixed refers to stories with contextual features whose valence is mixed. Neutral refers to stories with contextual features whose valence is neutral. Whiskers indicate one standard error of the mean.

Figure 2 illustrates that there is a pronounced story-statistic gap for both positive and negative reviews. Going further, we find no difference in recall by whether the reviews are positive or negative (*Story*: $p = 0.328$, *Statistic*: $p = 0.991$). Moreover, there is no heterogeneity in effects by valence of the quantitative information on change in belief impact for *Story* ($p = 0.860$). We observe, however, that positive statistics affect beliefs more persistently compared to negative statistics ($p < 0.001$).

Features of decoy information. In our baseline design, each respondent received one statistic and one story across the three scenarios. As a result, any given story was accompanied by a statistic as the "decoy" information in the respective other scenario, whereas any given statistic was accompanied by a story in the decoy scenario. To examine how sensitive the story-statistic gap is to the structure of decoy information, we systematically manipulated the number, type and valence of decoy pieces of information in a separate experiment (see Appendix Section A.1 for details). In a between-subject design, respondents either received two statistics, two stories or twice no information as decoys. We document a robust story-statistic gap across all conditions. In fact, the estimated story-statistic gap has a similar magnitude, irrespective of the number, type and valence of decoy information. This suggests that the story-statistic gap is robust to the basic structure of the decoys in our baseline setting.

2.4 Interpreting the Story-Statistic Gap

There are many ways in which stories differ from statistics, which begs questions about what exactly the story-statistic gap captures. We here provide a brief overview of several key differences and how they relate to the story-statistic gap as well as to our model-guided examination of mechanisms in the subsequent parts of this paper. Rather than an in-depth discussion, this section points out an array of considerations, some of which turn out to be less relevant for our purposes than one might expect, and foreshadows relations to our mechanism experiments.

Engagement with additional information and processing time. We view differences in processing time, which may be indicative of the encoding strength, as one plausible mechanism relating to 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.3 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. Note that the focus of our mechanism experiments will be on manipulations of cross-similarity, the similarity between a target scenario and other decoy scenarios, all of which hold the processing time of the target scenario constant.

Deliberation. A concern about beliefs formed in *Immediate* is that it might take some time for information to "sink in" to be fully processed. In that case, using the immediate belief as a benchmark may not adequately capture the maximal belief update. This would compromise inference about the story-statistic gap if such deliberation occurs to different degrees for stories and statistics. Given the nature and content of the scenarios and information, we see little reason for such prolonged deliberation to play much of a role. Empirically, we note that all of our key results on delayed versus immediate belief impact are supported by evidence on recall accuracy, which are not dependent on the immediate updating benchmark.

Emotions and vividness. Research in psychology has established a connection between emotions and memory (e.g., Kensinger and Schacter, 2008). Intuitively, stories tend to evoke emotions and are more vivid than statistics. First, our evidence on the valence of story content partly speaks directly to this mechanism. It suggests that while stories with more consistent qualitative features are recalled at somewhat higher rates than stories with mixed and neutral contextual features, these differences are relatively small, especially compared to the large differences in recall between stories and statis-

tics. Second, 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. Moreover, in the online samples we use, stories and statistics may be differentially surprising or typical given other studies they participate in. We fully embrace these issues and point out that they would also affect the response to stories versus statistics outside of our experiment. Moreover, we will examine mechanisms of cue-dependent memory that operate independently of baseline differences in the background memory database.

Mental representation of stories. An interesting question is how the story-statistic gap relates to how memories are coded and retrieved in the brain. Importantly, we explicitly do not claim that any of our evidence directly speaks to cognitive representations of memory in the brain, i.e., we do not wish to purport that stories are an elementary format of information storage in the brain. Instead, while we do not see our evidence as indicating that the brain encodes stories directly, we believe it suggests that stories *facilitate* efficient storage and retrieval of information. Relative to statistics, the story format is plausibly a facilitator of brain processes related to memory, such as mental imagery.

3 Outline of Conceptual Framework

We here outline a simple model of cue-dependent memory that adapts Bordalo et al. (2021a) to stories and statistics. The model allows us to derive formal predictions for the differential recall of stories and statistics and provides a guiding structure for the empirical analysis of underlying mechanisms. All details and formal derivations are relegated to Appendix G.

3.1 Setup

Agent i forms beliefs about the quality of different products $j \in \{1, ..., n\}$. For each product, there exists a number r_j of reviews, each of which is positive or negative. The true fraction of positive reviews is drawn independently and uniformly for each product. There are two periods $t \in \{1, 2\}$ that we may think of as "past" and "present." In both

periods, the agent needs to state a belief for each product, $b^t_{i,j}$, about the likelihood that a randomly selected review for this product is positive. Our key interest is in belief formation at $t = 2$. For simplicity, we will assume Bayesian belief formation given the information that is recalled, so that any deviation from a benchmark of perfect recall and normatively optimal belief updating originates from selective recall, rather than non-Bayesian belief formation.

In $t = 1$, the agent might receive additional information about a product. We distinguish two types of information, statistics and stories. A statistic for product j consists of n_j random draws (without replacement) from the total number r_j of existing reviews. Let m_j denote the number of positive random draws. A story for product j consists of one randomly drawn review from the total number of reviews r_j, plus anecdotal information that describes the reviewer's experience with the product.

3.2 The Structure of Memory

The decision-maker (DM) has a memory database containing two types of memories. The first type is quantitative and exact in nature. Because this idea of recalling explicit facts is related to the concept of semantic memory in psychology research, we will refer to these as *semantic* memories (Kahana, 2012). In our case, semantic memories correspond to the precise statistic that the DM received in a scenario or the objective information about the review rating (one positive or negative review) contained in a story. Because the DM forms Bayesian beliefs based on what is being recalled, recalling and updating based on semantic memory is equivalent to assuming that the DM correclty recalls their posterior belief formed at $t = 1$.

The second type are episodic memories, which are experiences rather than quantitative facts. An episodic memory trace is a vector of binary features, where 1 denotes that a certain feature is present and 0 that it is not. E describes the set of all episodic memory traces, formed both inside and outside of the experiment. For simplicity, we assume that each product scenario in our experiment creates one episodic memory trace (see Appendix H for an extension allowing for multiple traces). In our setting, each episodic memory contains context features that encode the name of the product and the context in which the memory was formed, e.g., the time and place of the experiment. A story trace additionally includes features for each element of the anecdotal information, as well as the valence of the review (positive versus negative).[14] A statistic trace in episodic memory only comprises quantitative information. Such quantitative information encoded in episodic memory is potentially subject to information loss. For example, the DM may only encode the equivalent of the story valence as an experience,

[14]For simplicity, we abstract away from the qualitative information stories might entail.

i.e., whether the statistic was overall positive or negative. However, episodic memory might also be highly accurate and preserve much or all of the quantitative information in a statistic. To what extent episodic memory leads to "gisting" of statistics is an open question that we empirically explore in Section 5. As a starting point, we will assume that the DM only retrieves the valence of the statistic, i.e., whether it had a majority of positive, a majority of negative or equally many positive and negative reviews. The DM does not retrieve the precise sample size and fraction of positive reviews. Given the total number of reviews for a product, the DM then forms a Bayesian update based on the expected weight and extremity of the statistic, conditional on its valence.[15]

Given this dual structure of memory, note that each scenario in our experiment creates at most two traces: an episodic one and, conditional on actually receiving information in a scenario, a quantitative one. Further note that the model yields three possible outcomes of the memory retrieval process. First, if semantic memory is involved, information is fully recovered. Second, episodic memory might lead to a retrieval failure due to interference. Third, episodic memory might yield successful recall, but the recalled information is gisted.

3.3 Recall and Similarity

We model episodic memory as cued recall, i.e., memory retrieval upon being presented with a cue. The cue is the task description, i.e., the prompt to state a belief $b_{i,j}$.

The DM samples once from their memory database. If they retrieve a relevant memory trace, i.e., one of those coded in the scenario matching the cue, the information is used to form a Bayesian update. If they retrieve an irrelevant memory trace, i.e., one that contains no information about the cued product, they discard it and state the prior.

Following Bordalo et al. (2021a), we assume that the DM correctly recalls the quantitative memory for a given scenario with probability $(1-p)$, but resorts to episodic memory with probability p. The probability p thus captures the reliance on experiences.[16]

Sampling from episodic memory E is shaped by similarity and interference. The symmetric function $S(e_1, e_2) : E \times E \to [0, 1]$ describes the similarity between two episodic memory traces. It increases in the number of features shared between e_1 and e_2 and reaches a maximum of 1 at $e_1 = e_2$.

Conditional on sampling from the episodic memory database E upon being presented with the cue c, the probability of recalling the target cue e_c is:

[15] Specifically, the DM knows that the sample size was randomly drawn, and further assumes that the fraction of positives they saw was also random conditional on the valence.

[16] While we here start with the assumption of identical p for stories and statistics, the model accommodates that p differs by information type.

$$r(e_c, c) = \frac{S(e_c, e_c)}{\sum_{e \in E} S(e, e_c)} = \frac{1}{\sum_{e \in E} S(e, e_c)} \tag{1}$$

The denominator of this equation captures interference: we cannot fully control which memories we retrieve and sometimes recall irrelevant traces. All traces in E compete for retrieval, and the likelihood of retrieving a given non-target trace increases in its similarity to the target trace. We refer to the similarity between the target trace and non-target traces as *cross-similarity*.

3.4 Predictions

Prediction 1. *There is a story-statistic gap in memory: the effect of a statistic on beliefs decays more rapidly with a delay than the effect of a story.*

The retrieval process formalized here has three potential outcomes with different signatures in the evolution of beliefs over time. First, the exact recall of the quantitative information or the previously stated posterior through semantic memory leads to zero decay of belief impact over time. Second, successful recall of the target experience through episodic memory may or may not lead to belief decay, which depends on the nature of the potential information loss in episodic encoding. Third, failure to retrieve either a semantic or episodic memory trace leads to full decay, i.e., beliefs fully revert to the prior.

Under the assumption of an exogenous likelihood p of relying on episodic memory (as in Bordalo et al. (2021a)) that is independent of the information type, our basic model accommodates a story-statistic gap in two ways. First, retrieval failure is more likely for statistics than for stories. This is due to statistics exhibiting higher cross-similarity with irrelevant traces. The model generically captures higher cross-similarity for a statistic because a story trace comporises additional entries that encode the qualitative information and make it more distinct from irrelevant traces. Second, conditional on successful retrieval from episodic memory, decay can still be more pronounced for a statistic than for a story. This occurs to the extent that gisting of statistics in episodic memory leads to a bigger information loss than gisting of stories. An instructive way to think of these two potential avenues for a story-statistic gap through episodic memory is, first, the "extensive margin" effect of failing to retrieve the right trace and, second, the "intensive margin" effect of a differential information loss when recalling the correct trace. We will examine these channels qualitatively in Section 4 and quantitatively by means of a decomposition exercise in Section 5.

Prediction 2. *Adding contextual features to a piece of information decreases belief decay.*

Contextual features are encoded as additional entries in episodic memory traces. This generally decreases cross-similarity to non-target traces and thereby improves the recall likelihood of the target trace, decreasing the decay of belief impact over time.

Prediction 3. *Differences in cross-similarity between stories and statistics drive the story-statistic gap.*

Cross-similarity can operate along various margins. Depending on this margin, increases in cross-similarity can have stronger effects on recall and delayed belief impact of stories than statistics. First, increasing the number of scenarios has a stronger interfering effect for a given target statistic than a story, because of the higher baseline distinctiveness of stories, i.e., the higher number of features encoded in a trace of stories compared to statistics. Second, a higher similarity between the content of different stories creates competition for retrieval and increases interference. Third, more similar cues, i.e., more similar scenarios, interfere with the association between specific cues and pieces of information and thereby impede recall of both stories and statistics.

3.5 Extensions

Our basic model aims for parsimony and abstracts from various features of cue-dependent memory that have been shown to matter in practice. We here outline how we accommodate these features in extensions of the model.

3.5.1 Self-similarity

Next to cross-similarity, models of similarity and interference sometimes allow for a role of *self-similarity* (see, e.g., Bordalo et al., 2021b). Intuitively, self-similarity measures how similar a target trace is to other target traces, which plays a role when there are multiple target traces. Our baseline version of the model abstracts from self-similarity and delivers sharp behavioral predictions generated by the role of cross-similarity alone. However, the numerator of equation (1) naturally accommodates variation in self-similarity: while we restrict our focus to a single target trace which is maximally self-similar by definition, in the presence of multiple target traces, denoted as set T, the numerator becomes $\sum_{e\in T}\sum_{a\in T} S(e,a)\frac{1}{||T||^2}$ and is generally smaller than one. Our investigation of mechanisms in the subsequent section will focus on cross-similarity but also examine self-similarity.

3.5.2 Alternative Modeling Approaches

We acknowledge that there are many different ways of modeling cue-dependent memory. Our objective is to provide what we think is a disciplined, parsimonious way of adapting existing theoretical work in a way that accommodates our distinction between stories and statistics. Plausible extensions include the idea that information, and specifically stories, create multiple traces in episodic memory and that people may sample more than once. In Appendix H, we derive similar predictions to the ones outlined above for a model with several more general features that broadly follows the setup of Bordalo et al. (2021b).

4 Mechanisms

Guided by the predictions spelled out in Section 3, we proceed with our analysis of mechanisms in three steps. First, in Section 4.1, we test Prediction 2 on the power of adding contextual features. Second, we delve into the features of cross-similarity and interference, motivated by Prediction 3, in Section 4.2. Finally, we extend our investigation of the key channels of cue-dependent memory by studying self-similarity in Section 4.3.

4.1 The Role of Contextual Associations

Design. To causally examine the role of adding contextual features, we prompted respondents to imagine a typical review for the statistic or for a single review they learn about. Note that this intervention does not provide any objective information, qualitative or quantitative, allowing us to identify the distinct effect of associating obviously fictional contextual features with a piece of information in memory.

We implemented 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 contextual features, we designed 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 contextual information. The *NoStoryPrompt* condition is identical to *NoStory* except that respondents that received information about a single review were 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.

To obtain an incentivized measure of recall instead of the open-ended measure from the baseline experiment, we implemented a structured recall task.[17] We asked respondents to indicate which type of information they received about a given product. We further asked 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.[18] Unless respondents indicated that they did not receive any information about this product, we additionally asked them to indicate whether the information they received was positive or negative.[19] Respondents were 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.[20] The pre-registration for this experiment is available on AsPredicted, see `https://aspredicted.org/v9gk7.pdf`.

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 that it actually induced subjects to imagine reviews and write them down. The median (mean) number of words subjects wrote to describe an imaginary typical review was 22 (23). The text responses indicate that the vast majority of subjects made an honest effort to describe a review, such as in the following excerpt form 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

[17] As before, we randomized the order of our measures of recall and the belief elicitation in the follow-up survey.

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

[19] We tailored the question wording for respondents according to whether they indicated having received a single review, multiple reviews or "don't know".

[20] 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.

> that the gameplay will have 3D style graphics while in reality, it had very old-school-style graphics [...].

For ease of exposition, Figure 4 pools respondents in *NoStoryPrompt* and *StatisticPrompt*, as well as the *NoStory* and *Baseline* conditions.[21] The top panel of Figure 4 shows results on belief impact, while the bottom panel displays results on recall.

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.2 reveals that the difference-in-difference (difference in slopes) is also statistically significant ($p < 0.05$).

These patterns for *Delay* beliefs are underscored by results on recall. The bottom panel of Figure 4 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.2 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.

Our second main result is given as follows:

Result 2. *The addition of contextual features causes a more pronounced belief impact in delay and facilitates more accurate recall of information.*

4.2 Cross-similarity

We present three experiments that jointly aim to examine the importance of cross-similarity in a comprehensive fashion. We investigate the role of (i) the number of product scenarios presented within the experiment, (ii) the similarity of different pieces of information, and (iii) the similarity of product cues.

4.2.1 The Number of Product Scenarios

A key prediction of our model is that increases in cross-similarity via a higher number of product scenarios tend to more strongly impede the recall of statistics than stories. The rationale for more muted effects of this variation in cross-similarity on stories is that the richness of anecdotal content makes stories distinct and hence less similar to additional product scenarios.

[21] Table A.2 shows results separately for all 4 conditions and confirms that the disaggregated results are similar.

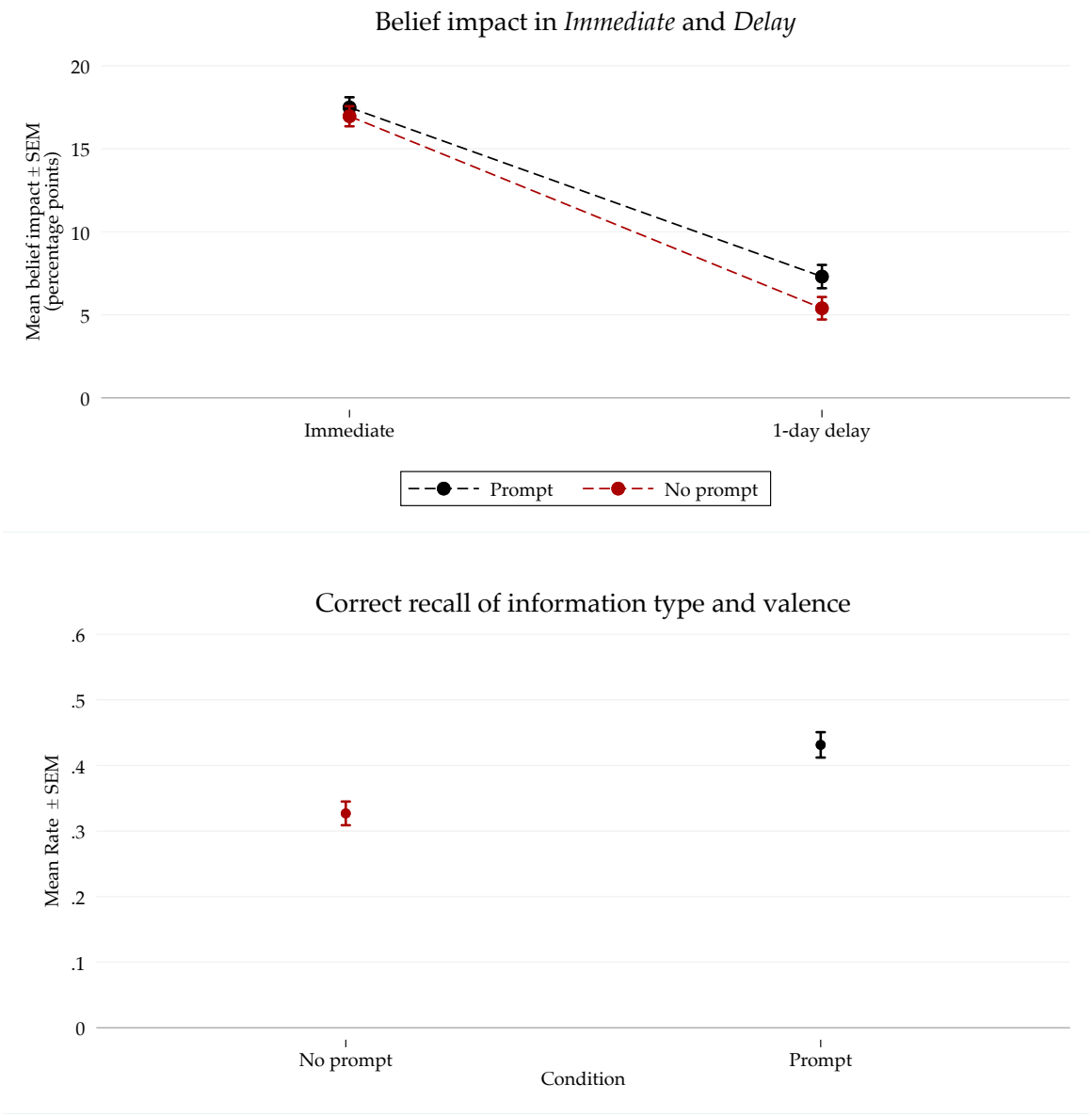


Figure 4: Belief impact and recall in Mechanism Experiment 1 (666 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and valence of information they received in the baseline survey. The red markers illustrate belief impact and recall for *No Prompt*, while the black markers illustrate belief impact and recall for *Prompt*. Whiskers indicate one standard error of the mean.

Design. The design broadly follows the structure of the main experiment. The key difference was that we varied, between-subject, whether there were one, three or six product scenarios. In the *1-product* treatment, there was 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 saw three scenarios and received two pieces of information, one story, one statistic and once no information. In

the *6-product* treatment, subjects saw six scenarios overall and also received 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 allowed us to cleanly study the effects of the number of product scenarios, while holding the total pieces of information constant.[22]

To keep incentives exactly constant between the different conditions, subjects in all treatments completed 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 completed 5 dot estimation tasks, while respondents in the 3-product treatment arm completed 3 dot estimation tasks, and respondents in the 6-product treatment only faced product-related tasks. The experimental instructions for the dot estimation task, in which subjects had to guess the number of dots displayed in a box for a short period of time, can be found on the following link: `https://raw.githubusercontent.com/cproth/papers/master/SSM_instructions.pdf`.

Sample and pre-registration. 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.[23]

The pre-registration for this experiment can be found on AsPredicted, see `https://aspredicted.org/as7i7.pdf`.

Prediction. The magnitude of the story-statistic gap both in the decay of belief impact as well as recall accuracy increases with the number of scenarios.

Results. Figure 5 and Table A.5 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

[22]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.

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

with our model, the story-statistic gap widens with the number of product scenarios.[24]

This pattern is strongly supported by the recall data, see the bottom panel of Figure 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.

Viewed through the lens of the model, these findings suggest that the differential effect of the number of product scenarios on stories versus statistics arises from differences in cross-similarity rather than memory load. The rationale for muted effects of cross-similarity on stories is that the richness of anecdotal content makes stories distinct and hence less similar to other product scenarios.

4.2.2 The Similarity of Story Content

One key difference between stories and statistics is that statistics are intrinsically more similar to one another than stories: intuitively, the numbers 73% and 82%, for example, are less distinctive than two highly contextualized and idiosyncratic stories about different products. This higher cross-similarity in turn increases interference. To study the role of cross-similarity, we conduct experiments with stories only. These experiments directly manipulate the similarity between a target story and decoy stories.

Design. We designed two treatments to study the role of story similarity. The incentives and basic setting were identical to our main experiment. Participants in both conditions learned about three products: a cafe, a restaurant, and a bar. Unlike in our main experiment, respondents received a story in each of the three scenarios. The target story in both conditions that our analysis focuses on was a positive review about the bar. The stories about the restaurant and the cafe were decoy stories and both featured a negative review. In the *Baseline* condition, the three stories were distinct and specific to each cue. The bar story described the interior of the bar, the restaurant story focused on food quality, while the cafe story was concerned with the service quality. In the *Story Similarity* condition, we kept the target story about the bar identical to *Baseline*, but increased 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 were still a cafe, a restaurant, and a bar, but all stories revolved around the interior design of the respective places. Thus, our treatments fixed the target story and only manipulated the similarity between the two decoy stories and the target story. All other design aspects were identical between the conditions. Appendix D.2 reproduces all stories that we used.

[24]The story-statistic gap in belief impact is close to zero for the 1-product scenario. This reflects that memory constraints are not binding in this setting.

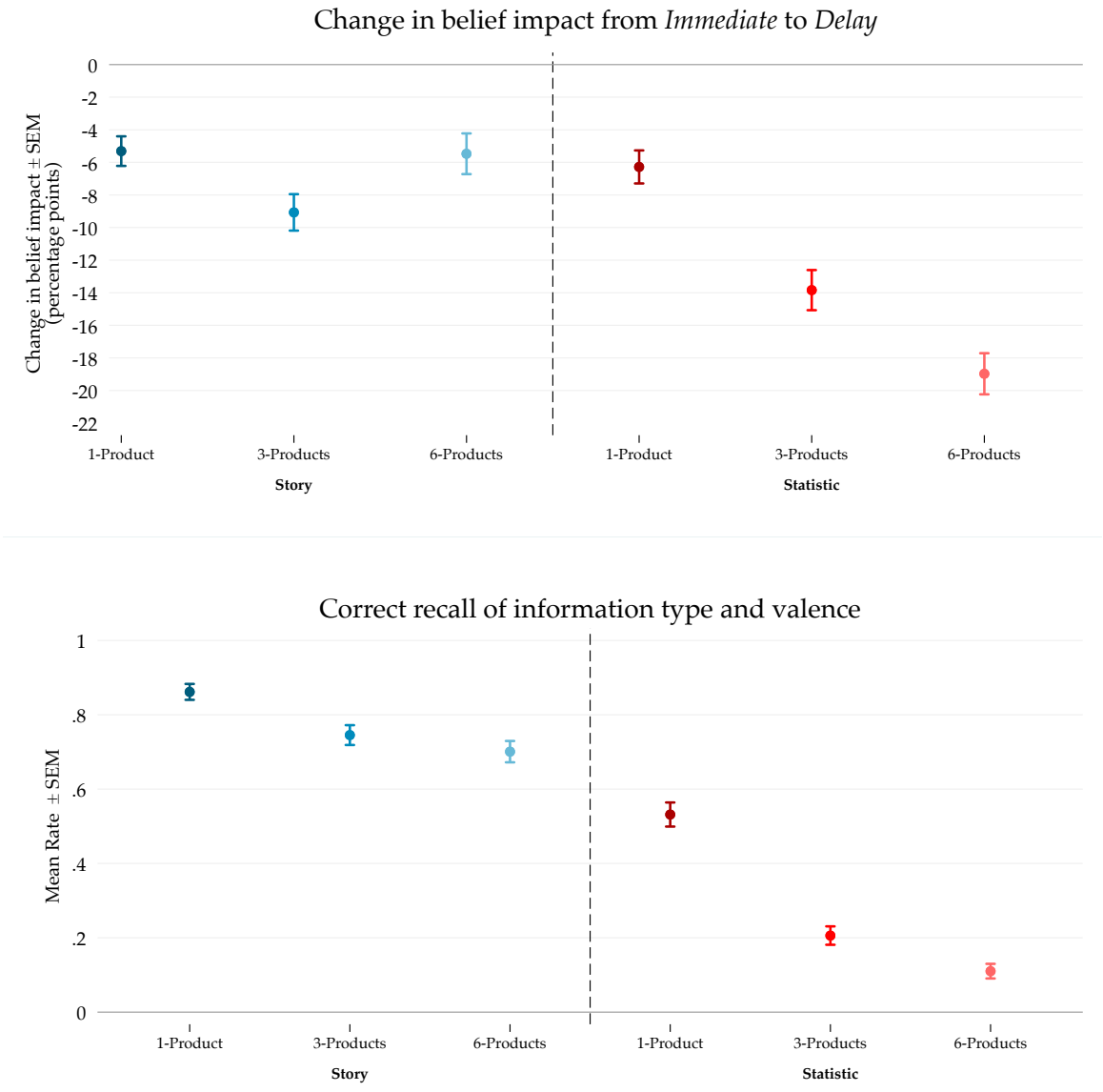


Figure 5: Change in belief impact and recall in Mechanism Experiment 2 (1,018 respondents). The top panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief 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 valence 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.

Sample and pre-registration. We recruited 1,150 respondents, of which 1,069 qualified for the follow-up. Respondents were randomized into the two conditions described above and a third condition described in Section 4.3.2. 879 respondents completed the follow-up survey. After the pre-specified sample restrictions, we have a sample size of

872, corresponding to a completion rate of 79 percent.[25] The pre-registration is available on AsPredicted, see `https://aspredicted.org/v7hh6.pdf`. The plan contains the two conditions described in this section as well as a third condition described in Section 4.3.2.

Prediction. The decay of belief impact is more pronounced in *Story Similarity* compared to the *Baseline* condition.

Results. The top panel of Figure 6 shows data on the belief impact of the target story in *Immediate* and *Delay*, separately for *Story Similarity* and *Baseline*. In line with the model prediction, the slope in belief impact is steeper in *Story Similarity* compared to *Baseline*. Delayed belief impact is significantly lower in *Story Similarity* than in *Baseline*, even though immediate belief impact is larger in the former condition. While average delayed belief impact in *Story Similarity* is 1.25 p.p. (s.e. 1.17), it is 4.43 p.p. (s.e. 1.09) in *Baseline*. Table 2 confirms this visual pattern and shows that the difference-in-difference 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 *Story Similarity*. This difference in 10 p.p. is statistically significant at conventional levels.

This finding has two 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 similar stories, retrieval of a target story gets crowded out and it becomes less likely that this story comes to mind.

4.2.3 Cue similarity

Our model posits that more similar cues should decrease delayed belief impact by increasing cross-similarity. Yet, our model remains silent about possibly differential effects for stories vs. statistics.

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. Subjects 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

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

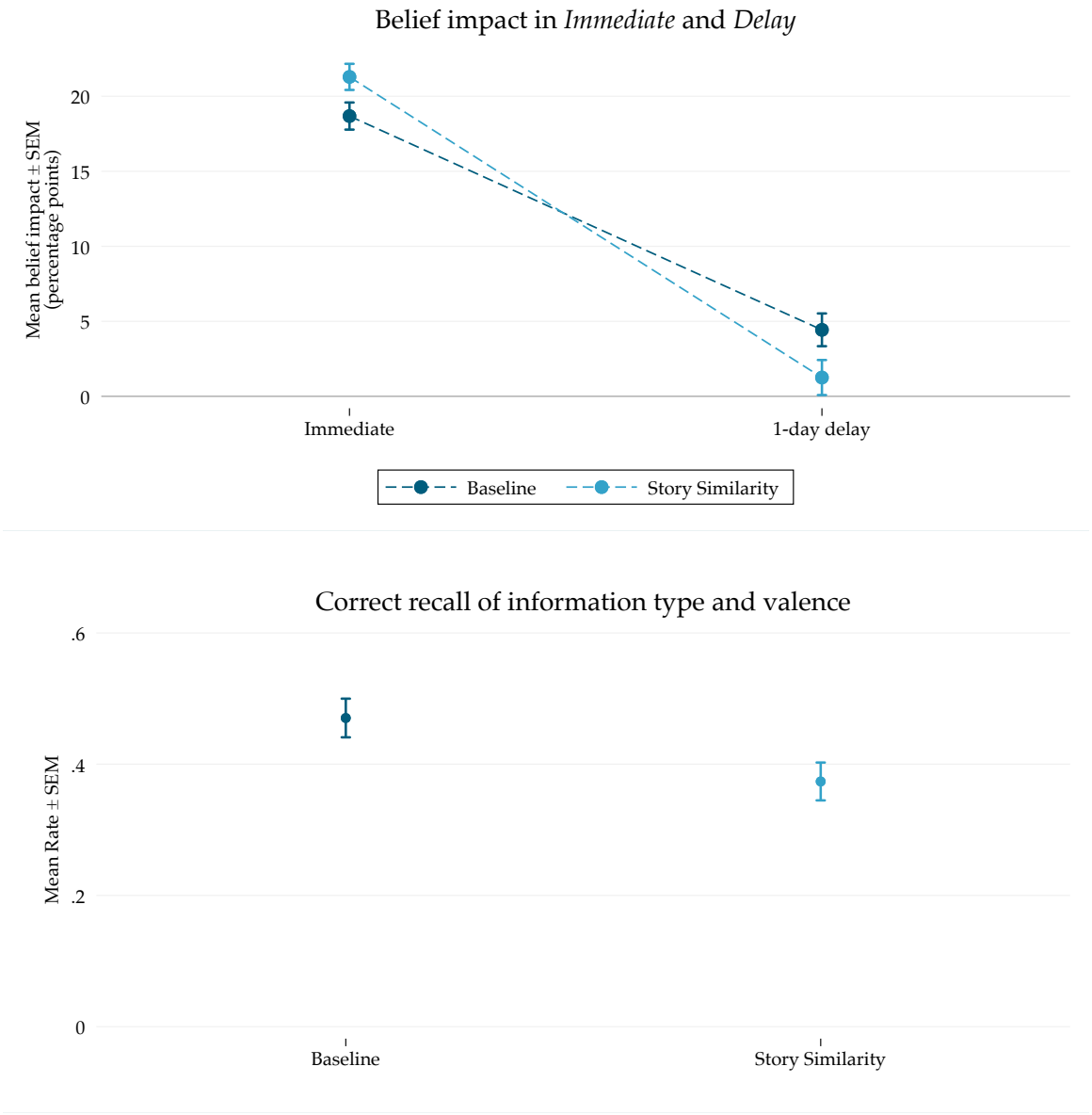


Figure 6: Belief impact and recall in Mechanism Experiment 3 (872 respondents). The top panel displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and valence of information they received in the baseline survey. The dark blue markers illustrate belief impact and recall for *Baseline*, while the light blue markers illustrate belief impact and recall for *Story Similarity*. Whiskers indicate one standard error of the mean.

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.[26] 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 7 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 7 largely displays the same pattern using our recall data. Table A.6 confirms this result.

Our third main result can be summarized as follows:

Result 3. *We report three experiments highlighting that cross-similarity significantly shapes delayed belief impact and recall. First, the story-statistic gap increases in the number of product scenarios. Second, delayed belief impact and recall of a story is impaired by higher similarity to stories in other scenarios. Third, an increase in the similarity of cues decreases belief impact in delay, both for stories and (somewhat less so) for statistics.*

4.3 Self-similarity

As outlined in Section 3, it is conceivable that the similarity between a cue and the information provided may shape recall and delayed belief impact. Appendix H provides a formalization that would yield this result on self-similarity as an extension to our baseline model. To test for the importance of self-similarity, we conduct experiments that manipulate (i) the similarity between a cue and the corresponding statistic and (ii) the similarity between a cue and the corresponding story.

4.3.1 Cue-Statistic Similarity

The similarity between a statistic and the cue might play a role to the extent that the *format* in which the statistic is provided resembles the format of the question that people are asked. For example, both might be presented in the similar format of a fraction, but can also be presented in less similar ways, as is the case in our main experiment, where

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

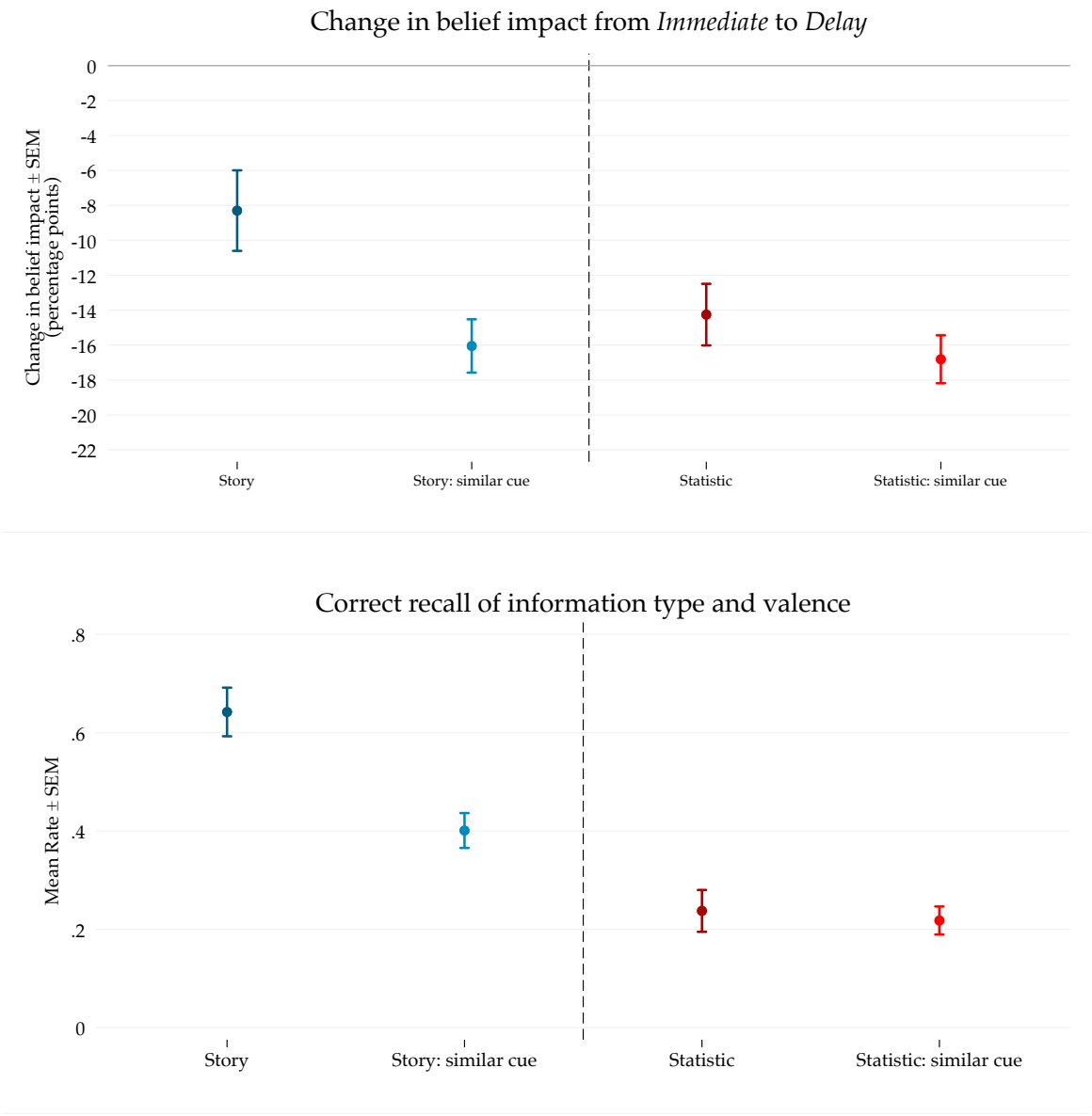


Figure 7: Change in belief impact and recall in Mechanism Experiment 4 (1,018 respondents). The top panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief 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 valence 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.

one is an absolute number and one a percentage. Put differently, self-similarity between the statistic and the cue might be driven by the question part of the cue.

Design. The experiment featured one key treatment variation. The *Dissimilar Format* treatment elicited beliefs as before – about the likelihood that a randomly chosen review is positive – and thus exactly corresponds to our main experiment. In the *Similar Format* condition, by contrast, we elicited beliefs about the percentage of positive reviews in the overall population of reviews of the product.[27] The rationale of this manipulation was that in *Similar Format*, the question people answered is more similar to the type of information they were provided with in the statistic condition, which is about the count of positive reviews in a subsample of reviews about the product.

As an additional similarity manipulation, we randomized whether the statistical information itself was expressed in terms of an absolute number of positive reviews in a subsample (*Statistic Dissimilar*) – as in our main study – or in terms of a percentage of positive reviews in a subsample (*Statistic Similar*). This means we ran a total of four between-subject conditions, reflecting a 2 (*Dissimilar Format / Similar Format*) × 2 (*Statistic Dissimilar / Statistic Similar*) factorial design.

The comparison between the *Similar Format* and the *Dissimilar Format* conditions allows us to examine how the similarity between the statistic and the cue affects recall. The *Statistic Similar* condition makes the additional information even more similar to the cue whenever the question format involves a fraction, creating a "high cue-statistic similarity" condition, providing us with additional variation to study the role of statistic-cue similarity.

Sample and pre-registration. 1,532 respondents completed the baseline survey and also met the inclusion criteria. 922 respondents completed the follow-up survey, corresponding to a 60 percent completion rate.[28] The pre-registration for this experiment is available on AsPredicted, see `https://aspredicted.org/ZFF_88V`.

Prediction. *Similar Format* and the interaction effect between *Similar Format* and *Similar Statistic* decrease the decay of belief impact and forgetting.

Results. Appendix Figure A.3 and Table A.4 document that the *Similar Format* has a positive, yet small effect on delayed belief impact and recall. The decay of belief impact is somewhat smaller in *Similar Format* than *Dissimilar Format*. This effect is more pronounced in the recall data, and reaches significance for the case of statistical information (column (4) of Table A.4), in line with the notion that a higher similarity of the question format to the statistic slightly improves retrieval.

[27] We accordingly adjust the description of incentives, which are framed in terms of guesses about the true percentage of positive reviews in this condition, but kept otherwise identical.

[28] The completion rate to the follow-up survey does not differ significantly across treatment groups ($p = 0.59$).

Moreover, we find that the effect of displaying statistical information as a percentage instead of an absolute number does not have significant effects on belief impact and recall. More specifically, we also do not observe a significant interaction effect between the question format and the display format of statistical information. A plausible interpretation is that these are already highly similar at baseline (in our main study), so that the manipulation of making them even more similar makes little difference. In practice, we might expect that the question format and the display format are much *less* similar to each other, and that variation in similarity across contexts plays a larger role.

Considering the insignificant effects on delayed belief impact and the mixed evidence on recall, we take this as, at best, suggestive evidence that the similarity of the question format to the piece of additional information importantly shapes forgetting.

4.3.2 Cue-Story Similarity

The qualitative information contained in stories is often intrinsically related to the corresponding cues, e.g., a story for the cue "Restaurant" will typically feature restaurant-related content. Stories are typically associated with the part of the cue that encodes the scenario name. As a result, the self-similarity of anecdotal information may be higher for stories than for statistics. To examine the role of self-similarity, we conduct experiments that manipulate the extent of similarity between stories and cues.

Design. We employed the same *Baseline* condition as in Section 4.2.2, but compared it to a different treatment, *Cue-Story Similarity*.[29] This condition relied on the same decoy stories for the cafe and the restaurant as *Baseline*. 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.

Prediction. Recall accuracy is lower in *Cue-Story Similarity* than in baseline.

Results. 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 2 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)

[29] Subjects were randomized within-session to either the *Cue-Story Similarity* condition, the *Story Similarity* condition or the *Baseline* condition

of Table 2 reports results on belief impact. Notably, 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 baseline levels of belief impact, but nevertheless highlights that the overall evidence of this manipulation for self-similarity is ambiguous. Our fourth main result is given as follows:

Result 4. *The effect of self-similarity for the story-statistic gap is, at best, of minor importance relative to the strong and consistent results of cross-similarity.*

Table 2: (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 Mechanism Experiment 3 (872 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. Column (3) and (4) display the fraction of respondents correctly recalling the type and valence of information they received in the baseline survey. OLS estimates, standard errors clustered at the respondent level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5 Decomposing the Story-Statistic Gap

The mechanism experiments reported in Section 4 provide causal evidence for the baseline model of cue-dependent memory outlined in Section 3 as an explanation of the story-statistic gap. In the following, we provide a heuristic decomposition of the gap into the different memory channels captured by the model. The purpose of this exercise is to move beyond experimental manipulations that establish the features of recall in a qualitative way and provide an approximate quantification of the different retrieval modes.[30]

To reiterate, the model accommodates three outcomes of the memory retrieval process, each possibly associated with a different signature in the decay of belief impact. First, the DM may resort to semantic memory and correctly recall all information (or, equivalently, their previously stated belief), which corresponds to a benchmark of zero belief decay (class *Exact Recall*). This occurs with exogenous probability $(1-p)$. Alternatively, with probability p, the DM relies on episodic experiences. The second outcome of cued recall is that the DM does not retrieve a target memory and therefore returns to the prior (class *Forgetting*). Such retrieval failure creates a second clear benchmark of full belief decay. Third, again conditional on relying on episodic memory, the DM successfully retrieves a target trace. In this case, the distinguishing feature is that the episodic memory trace of a scenario *may* be associated with information loss: in our baseline model, the DM may only remember the gist of a signal, i.e., its valence (positive or negative), but not its exact strength and weight (*Inexact Recall*). However, we stress that there is no guidance from previous empirical or theoretical work on the nature of such gisting of statistics and stories in episodic memory, and so it is possible that (almost) no precision is lost when information is successfully retrieved from episodic memory.[31]

The combination of recall and panel belief data allow us to shed some light on the relative magnitudes of these recall channels. The subsequent analysis focuses on our baseline experiment reported in Section 2.

First, the bottom panel of Figure 1 identifies the fraction of beliefs associated with imperfect recall of type and valence of information across conditions: Following this metric, the *Forgetting* class comprises 39 percent of observations in *Story*, but 73 percent

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

[31]The average signature of the potential information loss in belief decay is plausibly bounded by the other two classes: information loss through episodic memories should lead to *some* belief decay. These bounds seem plausible irrespective of the exact mechanisms underlying the information loss. If the information loss is associated with a form of valence gisting as in the baseline model, i.e., the DM treats statistical information based on the expected weight and extremity conditional on a statistic's valence, belief decay will fall between the two extremes of no and full decay established above. If information loss in episodic memory instead is associated with pure noise, i.e., a statistic is randomly retrieved as more or less extreme than the truth, then there will be no belief decay on average (conditional on successful retrieval).

in *Statistic*. According to the model, these observations reflect retrieval failure in episodic memory and should be associated with beliefs that fully return to the prior of 50%, or a corresponding belief impact of 0 in *Delay*. Figure 8 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 *Forgetting* indeed reverts to close to zero in *Delay*, as predicted by the model.

Second, among observations not classified as *Forgetting*, we may establish an upper bound for the class of *Exact Recall* that flawlessly retrieves relevant quantitative information from semantic memory: all cases in which beliefs stated in *Immediate* and *Delay* are exactly identical. This comprises 13.71 percent (22.47 percent of non-*Forgetting* observations) of all observations in *Story* and 6.03 percent (9.11 percent of non-*Forgetting* observations) of all observations in *Statistic*. Note that these figures only identify an upper bound because, in principle, successful retrieval from episodic memory might also be associated with no information loss, leading to the same pattern of no decay. This shows that even the upper bound for the class of exact, semantic memory corresponds to a relatively small share of observations.

Finally, we further zoom in on observations not classified as *Forgetting*. Again, this group associated with correct recall includes both the *Exact Recall* and *Inexact Recall* classes, and we know that a relatively small share of 22.47 percent in *Story* and 9.11 percent in *Statistic* exhibit zero decay. How large is the decay for the large share of cases where immediate and delayed beliefs differ, which we classify as *Inexact Recall*? Figure 8 reveals that there is zero average belief decay in the *Story* condition and a quantitatively minor, only marginally significant decay in the *Statistic* condition.

Taken together, the following main insights emerge: The lion's share of the story-statistic gap appears to be driven by the extensive margin of memory, i.e., differential retrieval failures for stories and statistics. This underscores our mechanism evidence on the central role of cross-similarity that drives interference. Next, a relatively small share of all observations ($<$ 15 percent) without retrieval failure corresponds to perfect stability of beliefs, i.e., even the upper bound for exact, semantic recall seems low, highlighting the importance of episodic memory. Finally, and perhaps most strikingly, this exercise clearly reveals almost no variation on the intensive margin of recall, i.e., there is close to no information loss when episodic memories are retrieved in our setting. Conditional on correct recall of the valence and type of information and on not stating exactly identical beliefs, we document virtually no story-statistic gap.

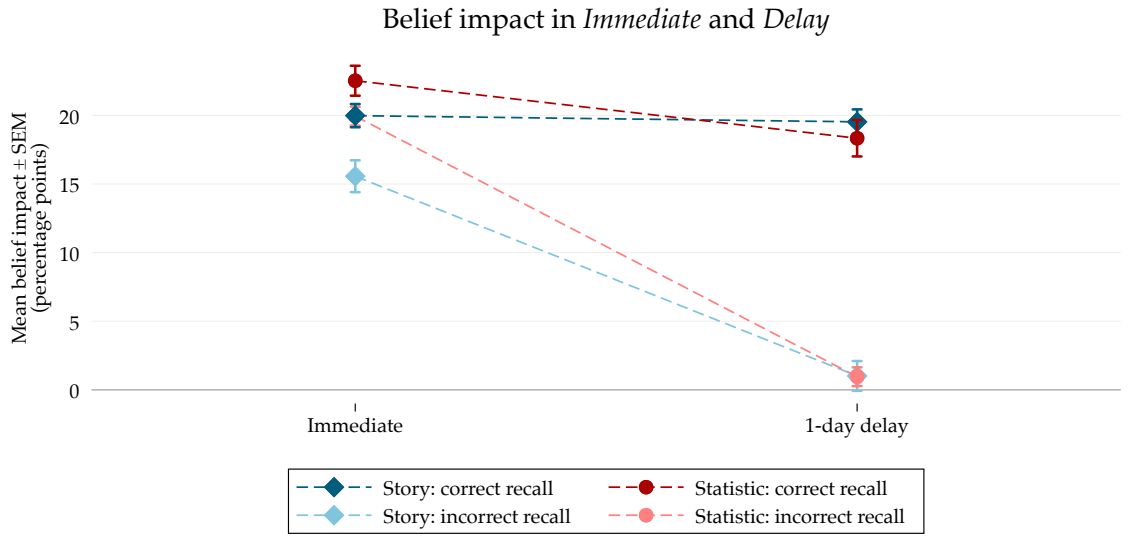


Figure 8: The decay of belief impact by recall accuracy in the baseline experiment (984 respondents). The figure displays belief impact in percentage points, separately for conditions *Immediate* and *Delay*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact 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 slower 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 rich contextual features of stories: adding context to statistics increases delayed belief impact and recall accuracy. Guided by a simple model of cue-dependent memory, we experimentally examine the explanatory power of different features of cross-similarity as sources of interference. Consistent with the model, our evidence suggests that similarity relationships are an important force behind the story-statistic gap. Stories tend to be distinct, whereas the abstract nature of statistics makes them similar to other, irrelevant statistics.

A key insight from our analysis of underlying mechanisms is that the features of memory that favor the recall of stories are not unique to stories. It does not seem to be the case that the way we store and retrieve stories is fundamentally different from how we store and retrieve statistics. Rather, cross-similarity and interference account for the lion's share of the story-statistics gap.

We establish two novel findings that inform future work on memory. First, the role of interference as driven by cross-similarity appears to be far more powerful than the effects of self-similarity in the settings studied here. Second, our memory decomposition

provides striking evidence that the extensive margin of successful retrieval from episodic memory is much more important for the persistence of information in beliefs than the intensive margin of gisting information.

A natural extension of our work is to examine *which* stories tend to be shared in practice. Conceivably, the most extreme and surprising stories are particularly likely to be told and re-told because they are "worth telling". If true, this would point to the possibly harmful implications of the story-statistic gap: the less representative the stories that are shared, the larger the final belief distortions, providing an explanation for the well-documented persistence of biased beliefs.

Our findings bear 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 contextual, anecdotal associations to ensure that the information sticks with the audience. For instance, statistical information about economic quantities should be coupled with anecdotal information that is consistent and inherently reminiscent of the embedded statistical information. Moreover, our results highlight 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.

References

Afrouzi, Hassan, Spencer Yongwook Kwon, Augustin Landier, Yueran Ma, and David Thesmar, "Overreaction in expectations: Evidence and theory," *Available at SSRN*, 2020.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva, "Immigration and redistribution," *Review of Economic Studies*, 2022.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective Models of the Macroeconomy: Evidence from experts and Representative Samples," *Review of Economic Studies*, 2022.

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

Bénabou, Roland, Armin Falk, and Jean Tirole, "Narratives, Imperatives, and Moral Reasoning," Working Paper 24798, National Bureau of Economic Research, July 2018.

Bordalo, Pedro, Giovanni Burro, Katie Coffman, Nicola Gennaioli, and Andrei Shleifer, "Imagining the Future: Memory, Simulation and Beliefs about Covid," *Working Paper*, 2021.

_ , John J Conlon, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer, "Memory and Probability," *Quarterly Journal of Economics*, 2021.

_ , 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," *The Quarterly journal of economics*, 2021.

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*, 2022.

_ , Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth, "Justifying Dissent," NBER Working Paper 27288 February 2022.

Charles, Constantin, "Memory and Trading," *Available at SSRN 3759444*, 2021.

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," *The Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

_ and Florian Zimmermann, "Correlation neglect in belief formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

_ , Frederik Schwerter, and Florian Zimmermann, "Associative memory and belief formation," Technical Report, National Bureau of Economic Research 2020.

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

Gennaioli, Nicola and Andrei Shleifer, "What comes to mind," *The Quarterly journal of economics*, 2010, *125* (4), 1399–1433.

Graeber, Thomas, "Inattentive inference," *Journal of the European Economic Association*, 2022.

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

Kahana, Michael Jacob, *Foundations of human memory*, OUP USA, 2012.

Kendall, Chad W and Constantin Charles, "Causal narratives," Technical Report, National Bureau of Economic Research 2022.

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

Kwon, Spencer Yongwook and Johnny Tang, "Extreme Events and Overreaction to News," *Available at SSRN 3724420*, 2020.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in contingent reasoning: The role of uncertainty," *American Economic Review*, 2019, *109* (10), 3437–74.

Michalopoulos, Stelios and Melanie Meng Xue, "Folklore," *The 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," *Available at SSRN 3919471*, 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.

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

_ , *Narrative economics*, Princeton University Press, 2020.

Online Appendix: Stories, Statistics and Memory

Thomas Graeber Christopher Roth Florian Zimmermann

A Additional Results

A.1 Robustness to Decoy Information

To further probe into the robustness of the story-statistic gap, we examine the role of features of the decoy information using an additional experiment, in which we systematically manipulate the type and valence of decoy information.

We exogenously manipulate the type of information for the two decoy scenarios. Respondents either received two statistics for the decoys, two stories or twice no information. In addition, in contrast to the baseline design, we fully randomize the valence of the information provided for each scenario.

In the follow-up survey, we elicit beliefs exactly as in the baseline survey.

Sample and pre-registration. 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.[1] The pre-registration for this experiment can be found on AsPredicted, see `https://aspredicted.org/qy3wq.pdf`.

Results. Figure A.1 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 decoy 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 decoy information. This is visible across both our beliefs data and the incentivized structured recall elicitation.[2] 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

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

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

the story-statistic 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.2 shows how belief impact and recall of stories vary depending on the valence of decoy information. Compared to the statistics benchmark, we again find a robust and sizable story-statistic gap across decoys of different valence. We further find that decoy valence has a small but directionally plausible effect on the size of the gap: when decoy information has the same valence as the target information, both recall and delayed belief impact is larger than when the decoy information is mixed or of opposite sign.

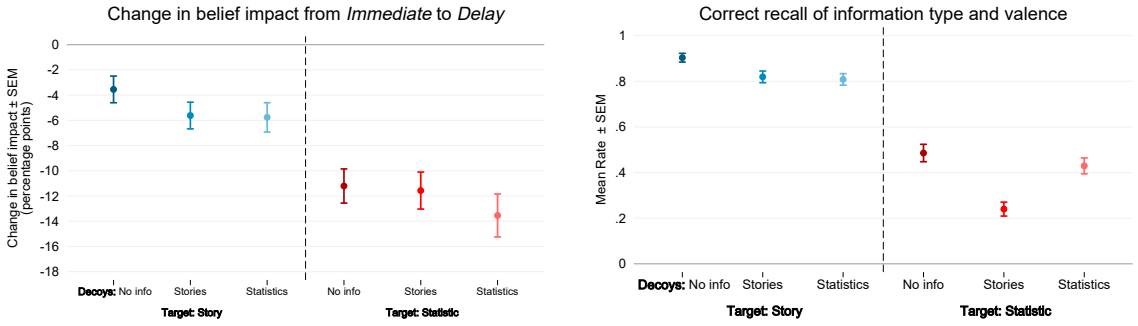


Figure A.1: Belief impact and recall in Robustness Experiment: The role of Decoy Information (1,513 respondents). The top panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between a stated belief and the prior (50%). Belief impact is signed in the direction of the rational update. The bottom panel displays the fraction of respondents correctly recalling the type and valence 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.

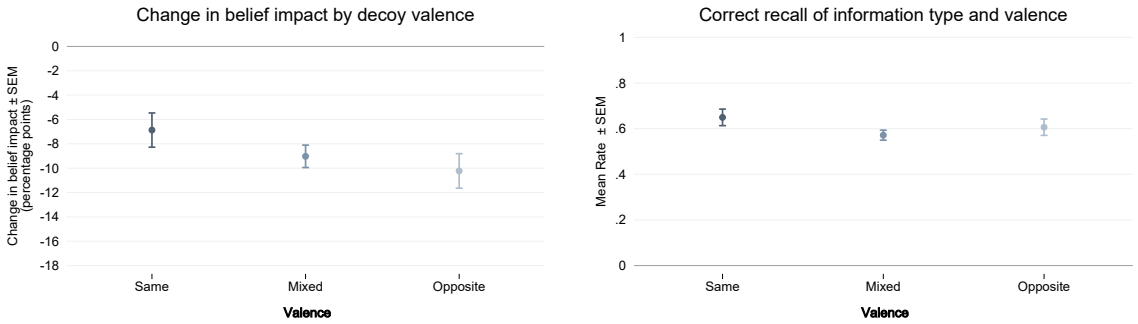


Figure A.2: Belief impact and recall in Robustness Experiment: The role of Decoy Information (1,513 respondents). The top panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance between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

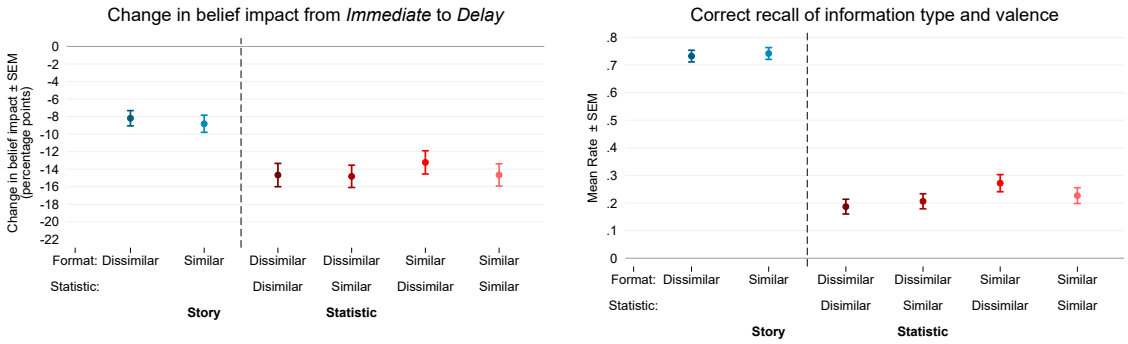


Figure A.3: Belief impact and recall in Mechanism Experiment 5: Question Format and statistic display (959 respondents). The top panel displays the change in belief impact in percentage points, defined as the difference in belief impact between *Delay* and *Immediate*. Belief impact is the signed distance 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 valence 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.

B Additional Tables

Table A.1: 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 contextual features: consistent, neutral, mixed.	Beliefs in immediate and delay; Open-ended recall in delay	https://aspredicted.org/e5mw7.pdf
Robustness Experiment: The role of Decoy Information	Prolific (1,513 respondents)	3 products (1 target and 2 decoy products): Target: Either Story or Statistic	Decoys: Either 2 stories, 2 statistics or 2 times no information	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/qy3wq.pdf
Mechanisms					
Mechanism Experiment 1: The role of associations	Prolific (666 respondents)	3 products. Decoys: Story and no information; Target varies across treatments	**Baseline condition**: statistic without prompt; **Prompt condition**: statistic with prompt; **No story condition**: Info on a single review without prompt; **No story prompt condition**: Info on a single review with prompt	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/v9gk7.pdf
Mechanism Experiment 2: 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
Mechanism Experiment 3: Story similarity and Cue-story similarity	Prolific (872 respondents)	3 products (bar, cafe and restaurant) with 3 stories	**Baseline**: 3 distinct stories about a bar, a restaurant and a cafe. **Story similarity**: same story about bar as in baseline, but now similar stories about a restaurant and bar. **Cue-story similarity**: As baseline, but the story about the bar is about an experience entirely unrelated and unspecific to a bar.	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/v7hh6.pdf
Mechanism Experiment 4: 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
Mechanism Experiment 5: Question Format and statistic display	Prolific (959 respondents)	3 products: story, statistic, no information	**Likelihood format**: same cue as in the baseline experiment. **Fraction format**: belief elicitation about the percentage of positive reviews **Statistic number display:** Statistical information is provided like in the baseline experiment, i.e. number of positive reviews. **Statistic percent display**: Statistical information is provided in terms of percentages.	Beliefs in immediate and delay; Structured recall task	https://aspredicted.org/ZFF_88V

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.2: Associations and contextual information: 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. Column (3) and (6) includes 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 valence of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: The story statistics gap: page time heterogeneity

	Dependent variable:			
	Belief Impact			Recall combined
Sample:	Immediate (1)	Delay (2)	Pooled (3)	Consistent (4)
Story	-2.46*	5.23***	-2.46*	0.33***
	(1.35)	(1.48)	(1.35)	(0.04)
Delay			-15.0***	
			(1.01)	
Story × Delay			7.69***	
			(1.55)	
Slow	0.43	0.26	0.43	0.091**
	(1.18)	(1.43)	(1.18)	(0.04)
Story × Slow	0.052	2.60	0.052	0.0013
	(1.84)	(2.15)	(1.84)	(0.05)
Delay × Slow			-0.16	
			(1.58)	
Story × Delay × Slow			2.55	
			(2.28)	
Control Mean	20.46	5.49	20.46	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 valence of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

	Dependent variable:			
	Belief Impact		Combined Recall	
Sample:	Story	Stat	Story	Stat
	(1)	(2)	(3)	(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 where 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) displays the fraction of respondents correctly recalling the type and valence of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: 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 valence of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 valence of information they received in the baseline survey. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Summary statistics

	Baseline Experiments		Mechanisms				
Experiment:	Baseline	Decoy	Association	Product	Story Sim	Cue Sim	Format
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Male	0.541	0.506	0.560	0.496	0.506	0.528	0.507
Age (years)	39.782	40.902	39.851	37.351	40.589	36.367	37.090
College	0.611	0.645	0.596	0.619	0.676	0.611	0.626
Employed	0.258	0.215	0.254	0.221	0.229	0.240	0.236
Observations	985	1,548	666	1,018	849	599	922

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): *Robustness Experiment: The role of Decoy Information*. Column (3): *Mechanism Experiment 1: The role of associations*. Column (4): *Mechanism Experiment 2: Number of product scenarios*. Column (5): *Mechanism Experiment 3: Story Similarity and Cue-story similarity*. Column (6): *Mechanism Experiment 4: Cue Similarity*. Column (7): *Mechanism Experiment 5: Question Format and statistic display*.

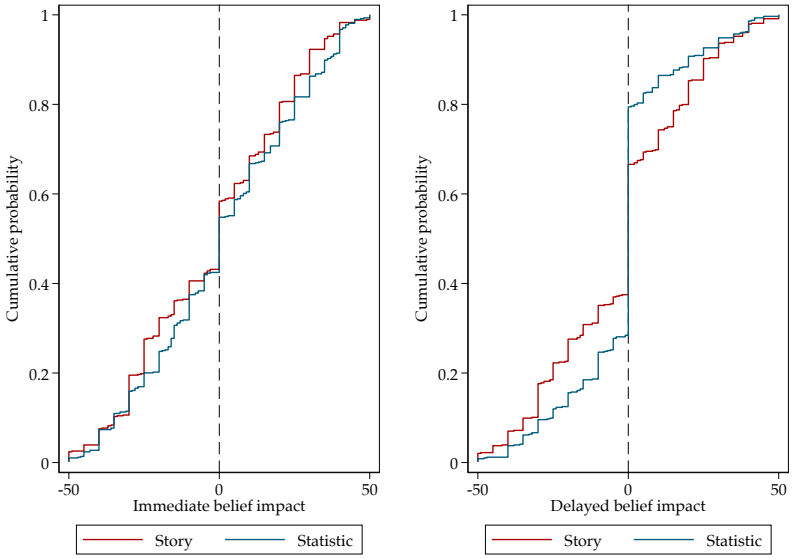
Table A.8: Attrition by conditions

	Dependent variable:						
	Wave 2 Completion						
Experiment:	Baseline	Decoy	Association	Product	Story Sim	Cue Sim	Format
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Neutral Story	0.012						
	(0.03)						
Mixed Story	0.020						
	(0.03)						
Decoy: Story		0.017					
		(0.02)					
Decoy: Statistic		-0.0054					
		(0.02)					
Prompt			0.0065				
			(0.05)				
1-Product				-0.014			
				(0.03)			
6-Products				-0.046			
				(0.03)			
Story Similarity					-0.017		
					(0.03)		
Cue Similarity					-0.019		
					(0.03)		
Similar Cue						0.020	
						(0.03)	
Belief: %							0.016
							(0.03)
Info: %							0.021
							(0.03)
Mean Completed	0.69	0.76	0.46	0.73	0.79	0.59	0.60
Observations	1437	2048	1442	1404	1069	1018	1532
p(Joint Null)	0.80	0.60	0.90	0.37	0.79	0.53	0.59
R^2	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): *Robustness Experiment: The role of Decoy Information*. Column (3): *Mechanism Experiment 1: The role of associations*. Column (4): *Mechanism Experiment 2: Number of product scenarios*. Column (5): *Mechanism Experiment 3: Story Similarity and Cue-story similarity*. Column (6): *Mechanism Experiment 4: Cue Similarity*. Column (7): *Mechanism Experiment 5: Question Format and statistic display*. The independent variables are indicators for each between-subject condition.

C Additional Figures

Figure A.4: 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.

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: Story similarity

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 Lindaâs 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 together 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 videogame, 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 valence, while the highest is given a “positive” valence. The product with the median fraction is assigned to the “no information” treatment, which doesn’t have a valence. 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 2.3) 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 valence 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 valence.

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.9: 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 subjects indicating âthey donât rememberâ 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 independent 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 talk 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 Conceptual Framework

In the following, we formally describe the conceptual framework outlined in Section 3 and derive the corresponding behavioral predictions.

G.1 Recall of episodic memories

Our framework distinguishes between exact recall of quantitative facts (or equivalently, ones previously formed posterior), labeled semantic memory, and recall of individual experiences, labeled episodic memory. Because we assume recall of semantic memories to be exact, we will mostly focus on the features of selective recall through episodic memory here.

G.1.1 Notation

Memory traces. Episodic memories are encoded as binary vectors indicating whether a certain feature is present or not. Every product scenario creates a single episodic memory e_j, where j indicates the product scenario.

The feature "Review-Experiment". Memories from different product scenarios are similar because they are part of the same experiment. The time and place the memories were made coincide. All memories are related to reviews and they may share additional structural elements, such as the display format of the additional information. For simplicity, we encode all the contextual features these memories share in common as a single feature called "Review-Experiment".

Cue. Participants are asked to assess the probability of a randomly drawn review of product j being positive. The cue is therefore given by "Review-Experiment"+"Product j". The cued set C_j consists of the single memory trace e_j which encodes the experience of this product scenario. In the extended version (Appendix H), the cued set C_j contains several memories.

Sets of memories. E will denote the set of all episodic memories in the memory database. Non-cued memories are given by the difference between all episodic memories and the cued memories, i.e. $E \setminus C_j = \overline{C}_j$. Other product scenarios are part of the non-cued memories. We denote the set of memories created during the experiment as R. Memories of non-cued product scenarios are therefore given by $R \setminus C_j = C_{-j}$ and memories from outside the experiment are given by $E \setminus R = \overline{R}$. We introduce a superscript to distinguish between the type of information given in the target and decoy product

scenario, where *story*, *stat* and *noinfo* represent the different types of information. For example, e_j^{story} represents the episodic memory trace of a story.

Recall. The probability to recall the target memory $C_j = \{e_j\}$ when being cued for exactly this memory is given by

$$r(C_j) = r(C_j, C_j) = \frac{S(e_j, e_j)}{\sum_{e \in E} S(e, e_j)} = \frac{1}{\sum_{e \in E} S(e, e_j)} \tag{2}$$

We can rewrite the denominator by splitting the episodic memories into different subsets:

$$r(C_j) = \frac{1}{1 + \sum_{e \in \overline{C}_j} S(e, C_j)} = \frac{1}{1 + |\overline{C}_j| \cdot S(\overline{C}_j, C_j)} \tag{3}$$

$$= \frac{1}{1 + |C_{-j}| \cdot S(C_{-j}, C_j) + |\overline{R}| \cdot S(\overline{R}, C_j)} \tag{4}$$

G.1.2 Assumptions on similarity

On the one hand, two memories become more similar when they share more features. On the other hand, two memories become less similar if there are more features by which to tell them apart, i.e., a feature is present in one trace but not the other. We now state the assumptions we will use to prove our predictions:

Assumption 1:

1. $S(C_j, C_{-j}) > 0$; $S(C_j, \overline{R}) > 0$
2. $S(C_j^{story}, C_{-j}^{story}) < S(C_j^{stat}, R \setminus C_j^{stat})$; $S(C_j^{story}, \overline{R}) < S(C_j^{stat}, \overline{R})$

The first assumption states that both for stories and statistics, the average similarity to other product scenarios and the average similarity to memories created outside the experiment is greater than zero. This is intuitive in the sense that there are always memories sharing some of the features, especially the ones from within the experiment. The second assumption states that the average similarity of a statistic (provided in the experiment) to memories within and outside the experiment is higher than that of a provided story to other memories. Intuitively, statistics are more generic than stories. Within the experiment, this is true because statistics are more similar to product scenarios without additional information, having less features by which to tell them apart. For memories outside the experiment, statistics are highly similar to most memories whereas stories are highly similar just to a few memories.

Assumption 1 implies that:

$$S(C_j, \overline{C}_j) > 0 \tag{5}$$

$$S(C_j^{story}, \overline{C_j^{story}}) < S(C_j^{stat}, \overline{C_j^{stat}}) \tag{6}$$

This is due to the following identity:

$$S(C_j, \overline{C}_j) = \frac{|C_{-j}|}{\overline{C_j}} \cdot S(C_{-j}, C_j) + \frac{|\overline{R}|}{\overline{C_j}} \cdot S(\overline{R}, C_j) \tag{7}$$

We assume the following ranking of similarities for different product scenarios, depending on the type of information given to the participants:

Assumption 2:

$$S(e_j^{stat}, e_{-j}^{stat}) > S(e_j^{stat}, e_{-j}^{noinfo}) > S(e_j^{stat}, e_{-j}^{story}) > S(e_j^{story}, e_{-j}^{noinfo}) \tag{8}$$

It is always the case that the products differ across scenarios. Statistics are highly similar to other statistics. The only difference between them is the product and the exact numbers of the statistic. Statistics are a bit less similar to scenarios without additional information since they do not have any statistical information. Statistics are most dissimilar to stories. Stories are longer, have a different structure and have more features that are not shared by statistics. A story is still more similar to a statistic than to not having any info since the two scenarios share the feature of having quantitative information. We didn't include the similarity of two stories in the ranking, since this heavily depends on the details of the story.

G.1.3 Proofs

Story vs. statistic:

Proposition 1 (Recall story vs. statistic)**.**

$$r(C_j^{story}) > r(C_j^{stat}) \tag{9}$$

Proof. The number of non-cued memories $|\overline{C}_j|$ does not change with the type of information of the target scenario. So the inequality directly follows from Assumption 1 b).

$$r(C_j^{story}) = \frac{1}{1 + |\overline{C_j^{story}}| \cdot S(\overline{C_j^{story}}, C_j^{story})} \tag{10}$$

$$> \frac{1}{1 + |\overline{C_j^{stat}}| \cdot S(\overline{C_j^{stat}}, C_j^{stat})} = r(C_j^{stat}) \tag{11}$$

□

The number of contextual features follows the same logic. Adding contextual features decreases cross-similarity and therefore increases recall by having lower interference.

Number of decoy scenarios: In the following, we assume that adding decoys does not change the average similarity of target to decoy. The reason is that the change in similarity depends on the type of decoy we add. Here we would like to make a more general statement. First, both for stories and statistics, the recall decreases with the number of decoy scenarios.

Proposition 2. *More decoy scenarios lead to a lower recall.*

Proof. Adding decoys leads to an increase in the memories created during the experiment, i.e., an increase in $|C_{-j}|$. Due to Assumption 1 a), we get the following inequality:

$$\frac{\delta r(C_j)}{\delta |C_{-j}|} = \frac{-S(C_{-j}, C_j)}{(1 + |C_{-j}| \cdot S(C_{-j}, C_j) + |\overline{R}| \cdot S(\overline{R}, C_j))^2} \tag{12}$$

$$= -S(C_{-j}, C_j) \cdot r(C_j)^2 < 0 \tag{13}$$

□

The effect is larger if the decoys are on average more similar to the target or if the recall before adding decoys is high.

Now we would like to compare the effect on stories and statistics. Statistics have a higher similarity to decoys, but stories have a higher initial recall. These two forces lead to an initially larger effect for statistics, but as we further increase the number of decoys eventually the effect will be larger for stories. We introduce x as a variable indicating the current number of decoys and $r_x(C_j)$ resembles the recall probability when having x decoys. Then we can make the following comparison:

$$\left|\frac{\delta r_x(C_j^{story})}{\delta x}\right| < \left|\frac{\delta r_x(C_j^{stat})}{\delta x}\right| \tag{14}$$

$$\iff \left(\frac{r_x(C_j^{story})}{r_x(C_j^{stat})}\right)^2 < \frac{S(R \setminus C_j^{stat}, C_j^{stat})}{S(R \setminus C_j^{story}, C_j^{story})} \tag{15}$$

Recall decreases when adding decoys. If the effect is initially larger for statistics, the gap between recall widens. So at some point the effect will be larger for stories. We summarize this in the following proposition:

Proposition 3. *Assume that (14) holds for $x = 0$. Then there exists an $\bar{x} > 0$, s.t. for all numbers of decoys $0 \leq x < \bar{x}$ the effect on recall when adding more decoys is greater for statistics than it is for stories and for all $x > \bar{x}$, the effect is greater for stories than statistics.*

Proof. Solving (14) for x leads to a quadratic function in x. The quadratic and linear coefficient are both strictly positive, given our initial assumptions. The intercept is strictly negative if (14) is fulfilled for $x = 0$. The proposition now follows. □

Similarity of decoys to target. If the similarity of decoys to target increases, the recall probability decreases.

Proposition 4. *Recall decreases with the similarity of decoys with target scenario.*

Proof.

$$\frac{\delta r(C_j)}{\delta S(C_{-j}, C_j)} = -|C_{-j}| \cdot r(C_j)^2 < 0 \tag{16}$$

Hence, making the decoys more similar decreases recall. □

We will next consider various features that make decoys more similar to the target scenario and therefore decrease recall. These include valence, type of decoy information, story similarity and cue similarity.

Changing decoys to have the same valence as the target scenario decreases recall.

Proof. Sharing the valence makes decoys more similar to the target scenario. So this directly follows from Proposition 4. □

Recall is lowest for the type of additional information with the highest similarity to the additional information of the target scenario .

Proof. This directly follows from Proposition 4. □

The previous means that statistics in the role of decoys lead to the highest interference for statistics as targets. We cannot make the same general statement for stories since the ordering depends on how similar the two products/stories are. In the baseline experiment, the two stories are least similar, leading to a higher interference when having a statistic as decoy.

Increasing the story similarity decreases recall.

Proof. This again increases similarity of target to decoy. Using Proposition 4 the claim directly follows. □

More similar cues lead to lower recall.

Proof. The only difference between the cues are the products. So increasing the cue similarity means to make the products more similar. Since the products are part of the episodic memory, this increases the similarity of the target memory and the decoys. We can again use Proposition 4 to complete the proof. □

G.2 Beliefs in time period $t = 2$

With probability $(1 - p)$, participants recall the quantitative information. This means they form the very same belief as in *Immediate*. With probability p, they rely on episodic memories. Given that participants recall the experience of the product scenario they are asked about, they are able to retrieve the valence as well as the type of information. Both for statistics and stories, they use the valence and, for simplicity, will update as having received a statistic of 1 or 0 out of 1 randomly drawn review being positive. We assume that a statistic has positive valence if the majority of reviews was positive.
We assume that participants sample once. There is no confusion, i.e., participants realize whether they retrieved the cued memory trace or not. If they recall a wrong memory, they do not update and state their prior belief.

G.2.1 Notation

Participants are asked to assess the probability of a randomly drawn review being positive. If the total number of reviews for a product j is r_j and the total number of positive reviews is m_j, the probability of sampling a positive review is $p_j = \frac{m_j}{r_j}$.

Let $b_{i,j}$ be the belief distribution over the total number of positive reviews M_j and $\hat{b}_{i,j}$ the belief distribution over the probability to draw a positive review. If we fix the

total number of reviews r_j, the two belief distributions have the following relationship:

$$\hat{b}_{i,j}(p) = b_{i,j}(p \cdot r_j) \tag{17}$$

This relationship allows us to focus on the belief distribution over the total number of positive reviews. We will denote the belief distribution over the total number of positive reviews of participant i in time period t by $b^t_{i,j}$. The time period can take on value 1 or 2. In both periods, participants can update their beliefs after potentially having received additional information. We will therefore use $b^0_{i,j}$ to denote the prior beliefs of participants before having received any additional information.

G.2.2 Prior belief

Participants are always informed about the total number of reviews r_j. The unknown variable is the total number of positive reviews M_j. Participants know that the number of positive reviews M_j is randomly drawn. This means that participants' prior beliefs follow a discrete uniform distribution with support $\{0, 1, \ldots, r_j\}$. This is equivalent to a beta-binomial distribution with parameters $\alpha = \beta = 1$.

$$M_j \sim BetaBin(r_j, 1, 1) \tag{18}$$

This yields the following density function:

$$b^0_{i,j}(m_j | r_j) = \frac{1}{r_j + 1} \text{ for } 0 \leq m_j \leq r_j \tag{19}$$

G.2.3 Posterior belief

There are three possible cases. Either the participant directly remembers the statistical information which happens with probability $(1-p)$, or the participant relies on episodic memories and retrieves the cued memory, which happens with probability $p \cdot r(C_j)$, or the participant relies on episodic memories and retrieves a non-cued memory, which happens with probability $p \cdot (1 - r(C_j))$. We will now derive the posterior for all cases.

First case A participant remembers the statistical information.

Proposition 5. *Let the total number of reviews for product j be r_j. If the statistical information of a scenario is given by k_j out of n_j reviews being positive, then a participant who remembers this information will state a belief of:*

$$\frac{(r_j - n_j)(k_j + 1)}{r_j \cdot (n_j + 2)} + \frac{k_j}{r_j} \tag{20}$$

Proof. The participant knows the total number of reviews r_j, as well as the sample size n_j and the number of successes k_j. The total number of positive reviews M_j is unknown, but participants form beliefs over the true value. As stated above, the prior is given by:

$$M_j \sim BetaBin(r_j, 1, 1) \tag{21}$$

Given that m_j is the total number of positive reviews and n_j reviews are drawn without replacement, the probability of having k_j positive reviews follows a hypergeometric distribution:

$$\pi(k_j | n_j, m_j, r_j) = \frac{\binom{m_j}{k_j} \cdot \binom{(r_j - m_j)}{(n_j - k_j)}}{\binom{r_j}{n_j}} \tag{22}$$

When updating according to Bayes' rule, the posterior belief over M_j is given by:

$$b^2_{i,j}(m_j | r_j, (n_j, k_j)) = \binom{r_j - n_j}{m_j - k_j} \cdot \frac{B(m_j + 1, r_j - m_j + 1)}{B(k_j + 1, n_j - k_j + 1)} \tag{23}$$

Here, $B(a, b)$ denotes the beta function. When defining $\alpha' = k_j + 1$, $\beta' = n_j - k_j + 1$, $r'_j = r_j - n_j$ and $m'_j = m_j - k_j$, one can see that the posterior follows a beta-binomial distribution:

$$M'_j \sim BetaBin(r'_j, \alpha', \beta'), \tag{24}$$

with support $\{0, \dots r'_j\}$. This is equivalent to $M_j \sim BetaBin(r'_j, \alpha', \beta') + k$ with support $\{k_j, \dots r_j - n_j + k_j\}$.

The payoff is maximized when reporting the mean of the belief distribution:

$$E[M_j] = E[M'_j] + k = r'_j \cdot \frac{\alpha'}{\alpha' + \beta'} + k \tag{25}$$

$$= \frac{(r_j - n_j) \cdot (k_j + 1)}{n_j + 2} + k_j \tag{26}$$

This leads to a reported probability of:

$$E[p] = \frac{E[M_j]}{r_j} = \frac{(r_j - n_j)(k_j + 1)}{r_j \cdot (n_j + 2)} + \frac{k_j}{r_j} \tag{27}$$

So the expected value lies in between $\frac{k_j}{r_j}$ and $\frac{(r_j - n_j + k_j)}{r_j}$, which are the only possible values after having observed k_j out of n_j successes. □

Second case: A participant retrieves the cued memory.

In this case the participant knows the total number of reviews as well as the type of information and the valence of the memory. For stories this is sufficient to update precisely. For statistics participants have to update conditional on knowing that (i) they received a sample of randomly drawn size and (ii) in line with the retrieved valence, more than half of the reviews were positive or negative.

Stories:

Proposition 6. *Assume that the participant recalls a cued memory and the additional information is a story. If the valence is positive, the participant will state a posterior belief of* $\frac{(2 \cdot r_j + 1)}{3 \cdot r_j}$ *and if the valence is negative, the participant will state a posterior belief of* $\frac{(r_j - 1)}{3 \cdot r_j}$

Proof. This directly follows from Proposition 5, when having $n_j = 1$ and $k_j = 0, 1$. □

Statistics:

Participants have to update their beliefs conditional on having received a positive or negative statistic. Let PV be defined as the set of statistics that lead to a positive valence:

$$PV := \{(k_j, n_j) | \quad \frac{n_j}{2} < k_j \leq n_j,\ 1 \leq n_j \leq r_j\} \tag{28}$$

We will now derive the probability participants report when they recall having received a positive statistic. The case of having received a negative statistic works analogously.

Proposition 7. *Assume that the participant recalls a cued memory and the additional information is a statistic. If the valence of the memory is positive, i.e.* $(k_j, n_j) \in PV$*, the participant will state a posterior belief of:*

$$\frac{\frac{1}{r_j} \cdot \sum_{(k_j, n_j) \in PV} \frac{(r_j - n_j) \cdot (k_j + 1) + k_j (n_j + 2)}{(n_j + 1) \cdot (n_j + 2)}}{\sum_{(k_j, n_j) \in PV} \frac{1}{n_j + 1}} \tag{29}$$

Proof. The posterior belief $b^2_{i,j}(m_j|r_j, PV)$ for $m_j \in \{0, \ldots, r_j\}$ is given by:

$$b^2_{i,j}(m_j|r_j, PV) = \frac{\pi(PV|m_j, r_j) \cdot b^0_{i,j}(m_j|r_j)}{\sum_{m_j=0}^{r_j} \pi(PV|m_j, r_j) \cdot b^0_{i,j}(m_j|r_j)} \tag{30}$$

$$= \frac{\pi(PV|m_j, r_j)}{\sum_{m_j=0}^{r_j} \pi(PV|m_j, r_j)} \tag{31}$$

$$= \frac{\sum_{(k_j,n_j)\in PV} \pi(k_j|m_j, r_j, n_j) \cdot \pi(n_j|m_j, r_j)}{\sum_{m_j=0}^{r_j} \sum_{(k_j,n_j)\in PV} \pi(k_j|m_j, r_j, n_j) \cdot \pi(n_j|m_j, r_j)} \tag{32}$$

$$= \frac{\sum_{(k_j,n_j)\in PV} \pi(k_j|m_j, r_j, n_j)}{\sum_{m_j=0}^{r_j} \sum_{(k_j,n_j)\in PV} \pi(k_j|m_j, r_j, n_j)} \tag{33}$$

For the last step we used that the size of the sample n_j is randomly drawn. This means that $\pi(n_j|m_j, r_j) = \frac{1}{r_j}$ for all $n_j \in \{1, \ldots, r_j\}$. Now we can state the conditional expected value $E[m_j|PV, r_j]$:

$$E[m_j|PV, r_j] = \frac{\sum_{m_j=0}^{r_j} m_j \cdot \sum_{(k_j,n_j)\in PV} \pi(k_j|m_j, r_j, n_j)}{\sum_{m_j=0}^{r_j} \sum_{(k_j,n_j)\in PV} \pi(k_j|m_j, r_j, n_j)} \tag{34}$$

$$= \frac{\sum_{(k_j,n_j)\in PV} \sum_{m_j=0}^{r_j} m_j \cdot \pi(k_j|m_j, r_j, n_j)}{\sum_{(k_j,n_j)\in PV} \sum_{m_j=0}^{r_j} \pi(k_j|m_j, r_j, n_j)} \tag{35}$$

$$= \frac{\sum_{(k_j,n_j)\in PV} \frac{(r_j-n_j)\cdot(k_j+1)+k_j(n_j+2)}{(n_j+1)\cdot(n_j+2)}}{\sum_{(k_j,n_j)\in PV} \frac{1}{n_j+1}} \tag{36}$$

The last step follows by using a version of the Chu Vandermonde identity and the expected value $E[m_j|k_j, n_j, r_j]$ we derived in Proposition 5. Multiplying the expression with $\frac{1}{r_j}$ leads to the stated probability that a randomly drawn review is positive. □

Third case: A participant retrieves a non-cued memory.

In this case, the participant has no information, except the total number of reviews r_j. The participant therefore relies on the prior.

$$M_j \sim BetaBin(r_j, 1, 1) \tag{37}$$

Again the participant states the probability maximizing the payoff, which is given by the mean, i.e., $\frac{1}{2}$.

Comparison of Updating when having a story vs. a statistic. If a story is given as additional information, participants state the same belief as in *Immediate* with probability $(1-p)+p\cdot r(C_j)$. With probability $p\cdot(1-r(C_j^{story}))$, they state the prior of $\frac{1}{2}$.

If a statistic is given as additional information, participants state the same belief as in *Immediate* with probability $(1-p)$. With probability $p\cdot r(C_j^{stat})$, they update in the right direction, but with a potentially lower intensity. With probability $p\cdot(1-r(C_j^{stat}))$, they state the prior of $\frac{1}{2}$.

H Formal Memory Framework: Extension

In the following we provide details on a model extension, accounting for self-similarity of cued memories.

H.1 Changes in extended version.

We extend the model by assuming that every product scenario creates several episodic memories. None of these memories contain the exact quantitative information but rather the valence of the experience. Since we allow for several memories attached to a single product scenario, self-similarity of these memories becomes important.

Memories of product scenarios A product scenario can be split in three parts, each consisting of a single or several memories. The first part introduces the product and the total number of reviews. The second part consists of the additional information. For a statistic this is a single memory. For a story there are several memories. One for the quantitative part and several memories encoding the anecdotal information. The third part consists of the participants immediate guess.

Cued set As in the basic version, the cue is composed of the context of the task and the product name, i.e. 'Context-Experiment'+'Name product'. The difference is, that the cued set now contains several memories, more specifically all memories created in the target scenario.

Recall The probability to retrieve a cued memory is given as the sum over the probabilities of recalling a memory belonging to the cued set of memories:

$$r(C_j) = \sum_{e \in C_j} r(e, C_j) \tag{38}$$

$$= \frac{S(C_j, C_j) \cdot |C_j|}{S(C_j, C_j) \cdot |C_j| + |\overline{C}_j| \cdot S(\overline{C}_j, C_j)} \tag{39}$$

$$= \frac{S(C_j, C_j) \cdot |C_j|}{S(C_j, C_j) \cdot |C_j| + |C_{-j}| \cdot S(C_{-j}, C_j) + |\overline{R}| \cdot S(\overline{R}, C_j)} \tag{40}$$

H.2 Proofs

The qualitative results from the short version of the model do not change, but we can make some additional predictions, when allowing self-similarity to play a role.

Story vs. Statistic In this section we compare the self-similarity of memories belonging to the same product scenario, when having a story vs. a statistic as additional information. Memories created in a scenario with a story C^{story} can be split in a part similar to the one when having a statistic C^{stat}, i.e. all memories not encoding the anecdotal information of the review and a qualitative part C^{qual}, i.e. additional memory traces encoding the details of the review only present when having a story. This can be summarized by $C^{\text{story}} = C^{\text{stat}} \dot{\cup} C^{\text{qual}}$.

Assumption 3:

1. $S(C^{\text{qual}}, C^{\text{qual}}) > S(C^{\text{stat}}, C^{\text{stat}})$
2. $S(C^{\text{qual}}, C^{\text{stat}}) > S(C^{\text{stat}}, C^{\text{stat}})$

The similarity of memories within a scenario with a statistic is only based on the two features 'Context experiment' and 'Name Product'.
Assumption a) can be justified, because the memories of the qualitative part in addition share specific details of the story. The memories have the same valence, are part of the same review, and share features of the specific experience.
Assumption b) can be justified, because the qualitative part in addition matches the valence of the memory encoding the quantitative information. Additionally the qualitative part is related to the product not only via context but also via content, making it more similar to the first part of the product scenario.

Proposition 8 (Average similarity story vs. statistic)**.**

$$S(C^{story}, C^{story}) > S(C^{stat}, C^{stat}) \tag{41}$$

This just means that the average similarity of two cued traces in a product scenario with a story is higher than it is in a scenario with a statistic.

Proof. Let $C = C_0 \dot{\cup} C_1$ and $S(C_1, C_1) > S(C_0, C_0)$ and $S(C_1, C_0) > S(C_0, C_0)$. Then we get the following inequality

$$\begin{aligned} S(C,C) \cdot |C| \cdot |C| &= S(C_0,C_0) \cdot |C_0| \cdot |C_0| + 2 \cdot S(C_0,C_1) \cdot |C_0| \cdot |C_1| + S(C_1,C_1) \cdot |C_1| \cdot |C_1| \\ &> S(C_0,C_0) \cdot (|C_0| \cdot |C_0| + 2 \cdot |C_0| \cdot |C_1| + |C_1| \cdot |C_1|) \\ &= S(C_0,C_0) \cdot |C| \cdot |C| \end{aligned}$$

It directly follows that $S(C,C) > S(C_0,C_0)$. □

Since a story consists of a statistical part as well as a qualitative part ,i.e. $C^{\text{story}} = C_{\text{stat}} \dot{\cup} C_{\text{qual}}$, stories create more memories than statistics:

$$|C^{\text{story}}| > |C^{\text{stat}}| \tag{42}$$

Proposition 1 together with (42) lead to the following:

[Self-similarity story vs. statistic]

$$S(C^{\text{story}}, C^{\text{story}}) \cdot |C^{\text{story}}| > S(C^{\text{stat}}, C^{\text{stat}}) \cdot |C^{\text{stat}}| \tag{43}$$

In the extended model, stories have an advantage over statistics in recall not only via cross-similarity, but also via self-similarity of the memories within the target scenario. So in the extended version the advantage in recall is even higher:

Proposition 9 (Recall story vs. statistic)**.**

$$r(C^{story}) > r(C^{stat}) \tag{44}$$

Proof. Using the analogous proposition in the short version together with Corollary H.2 directly proofs the result. □

Prompting Contextual Features

Proposition 10. *Adding memories to the target scenario increases recall.*

Proof. Adding memories to the target scenario means increasing $|C_j|$.

$$\frac{\delta r(C_j)}{\delta |C_j|} = \frac{S(C_j, C_j) \cdot |\overline{C}_j| \cdot S(\overline{C}_j, C_j)}{(S(C_j, C_j) \cdot |C_j| + |\overline{C}_j| \cdot S(\overline{C}_j, C_j))^2} > 0 \tag{45}$$

So the probability to recall a cued memory is higher if the number of cued memories increases. □

Cue-story Similarity

Proposition 11. *Increasing cue-story similarity increases recall*

Proof. Increasing Cue-story similarity means to make the anecdotal information of the review more related to the product itself. This means it increases the similarity of the memory traces belonging to the review to the memory trace encoding the title. This increases the average self-similarity $S(C_j, C_j)$.

$$\frac{\delta r(C_j)}{\delta S(C_j, C_j)} = \frac{|C_j| \cdot |\overline{C}_j| \cdot S(\overline{C}_j, C_j)}{(S(C_j, C_j) \cdot |C_j| + |\overline{C}_j| \cdot S(\overline{C}_j, C_j))^2} > 0 \tag{46}$$

Increasing self-similarity increases recall. □