From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Neuroscience

2017

Influence of contextual change on time perception, probability learning, and decision-making dynamics

Lositsky, Olga

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01sn00b141m

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Influence of Contextual Change on Time Perception, Probability Learning, and Decision-Making Dynamics

Olga Lositsky

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Program in
Neuroscience

Advisers: Kenneth A. Norman and Jonathan D. Cohen

September 2017

© COPYRIGHT BY OLGA LOSITSKY, 2017. ALL RIGHTS RESERVED.

Abstract

The principles that govern how memories are organized into categories, and how these categories then inform predictions for decision-making, constitute active areas of research. In this work, we investigated how people learn associations with different contexts (external cues that do not require responses but guide future behavior), how these associations are used to prepare decisions and how they influence time perception. The first study tested the hypothesis that people would interpret contextual changes in a story (including changes in spatial, social or temporal factors) as 'event boundaries', causing them to segment the story into distinct episodes. We predicted that more segmented intervals would appear longer in retrospect. Indeed, people's retrospective duration estimates were highly correlated with the number of event boundaries in an interval. Using fMRI, we found that activity patterns in brain regions important for binding spatial and temporal elements into episodes changed more during intervals with overestimated durations. The second study investigated how people learn when contextual cues signal a change in the situation. We hypothesized that people merge predictions across contexts into a single cluster when they are similar enough, and only form separate predictions for each context when truly necessary, as predicted by Bayesian inference models. We manipulated the degree to which cues predicted similar response or stimulus probabilities and built a new drift diffusion model to measure how people's expectations changed by cue. We found that people combined response and stimulus probabilities across cues when the probabilities were similar, despite maintaining entirely separate behavioral rules (stimulus-response associations) for each cue. Moreover, given the distortions in people's

probability memories (caused by merging), model fits showed that contextual predictions were used almost optimally in decisions. The third study tested whether contextual associations were always used to prepare the decision proactively, or whether they could inform the decision reactively. Simulations from the diffusion model reveal that across-trial probability learning, as well as working memory failures, must be quantified in order to isolate reactive strategies in behavior, and that response times provide a cleaner index of strategy than error rates.

Table of Contents

Abstract ..... **iii**

Acknowledgements ..... **x**

General Introduction **1**

Retrospective Time Perception and the Contextual Change Hypothesis *1*

Novelty of Study 1 with Respect to Existing Literature *8*

Hypothesis in Study 1 ..... *10*

Main Findings of Study 1 ..... *11*

Relationship between Study 1 and Study 2 ..... *13*

Novelty of Study 2 with Respect to Existing Literature *17*

Findings and Contributions of Study 2 ..... *20*

Relationship between Study 2 and Study 3 ..... *25*

References 27

Study 1: Neural Pattern Change During Encoding of a Narrative Predicts Retrospective Duration Estimates **33**

Abstract 33

Introduction 34

Results 39

Behavioral Results *39*

fMRI Results *48*

Anatomical ROI Analyses *52*

Whole-Brain Searchlights *61*

Comparing Results from ROI and Searchlight Analyses *64*

Patterns of activity in entorhinal cortex change slowly over time *66*

Story position effects cannot explain the correlation between duration estimates and neural pattern change *70*

Replication of Jenkins and Ranganath 2010: activity at encoding predicts accuracy of temporal context memory *72*

Discussion 75

Conclusion *83*

Methods 84

Participants *84*

Experimental Design and Stimuli *85*

Behavioral Data Analysis *88*

MRI Acquisition *91*

fMRI Data Preprocessing *92*

fMRI Data Analysis *99*

Acknowledgments 112

Prior Presentations, Publications and Co-Authorship 112

References 113

Study 2: Weak Fusion of Cue Predictions in Context-Based Decision-Making 122

Introduction 122

Background *122*

Our approach *127*

Methods Summary 130

Summary of Experimental Manipulation *130*

Participant Populations *132*

Implicit and Explicit Measures of Response Priors *134*

Model Predictions under Different Strategies *135*

Results 140

Outline of Results *140*

Experiment 1: Changing Trial Frequencies Across Participants *143*

Implicit Response Priors *143*

Explicit Response Priors *148*

Correlation between A and B Priors Varies by Similarity Between Cue Predictions *151*

Experiment 2: Changing Trial Frequencies Within Participants *160*

Relating Explicit and Implicit Measures: Do distortions in probability estimates account for sub-optimality in decision-making? *182*

Implicit and Explicit Priors are Closer to 0 than Optimal *193*

Alternative Interpretations and Additional Analyses *197*

Discussion 219

Conclusion *242*

Prior Presentations, Publications and Co-Authorship 244

References 245

Study 3: Distinguishing Proactive and Reactive Control Strategies in Behavior using a Drift Diffusion Model 249

Abstract 249

Introduction 251

Methods 253

The AX-CPT Task 253

Behavioral Experiments to Induce Proactive and Reactive Strategies 254

Drift Diffusion Model of Context-Dependent Decision-Making on the AX-CPT 254

Fitting the Model to Data 258

Landscaping Analysis: Finding where the models make distinct predictions 259

Results 261

Model Simulations 261

Behavioral Results 266

Model Goodness of Fit by Experimental Condition and Model Type 267

Conclusion 272

Acknowledgements 273

Prior Presentations, Publications and Co-Authorship 273

References 274

General Discussion 276

Contributions of Study 1 276

How findings from Study 1 inform Study 2: Do duration estimates increase when people "split" representations by context? 279

Contributions of Study 3 ..... *282*

How findings from Study 2 inform Study 3 ..... *285*

Conclusion 287

References 289

Acknowledgements

As students, we always act as sponges, absorbing the influences all around us. Nowhere has this metaphor felt more fitting than at the Princeton Neuroscience Institute, whose friendly, collaborative and supportive environment gave birth to the ideas in this thesis.

I do not even know how to begin to thank my advisers, Ken Norman and Jon Cohen, for the incredible amount of support and inspiration they have provided over the years. Their energy and passion for science have been infectious and I cannot even come close to emulating their rigor. Their encouragement was invaluable during the middle years of graduate school, which I found to be the most taxing. It was during those years that I developed sufficient mastery of the literature and methods to grasp how little we understood about the mind and brain – to see how many effects were weaker than they appeared, how few hypotheses could be confirmed with certainty, how many flaws hid within each hallmark experiment. Yet if it had not been for the support of my advisers, as well as my colleagues and friends in the department, I might never have passed to the next stage – the understanding that it is precisely these challenges that make science an endeavor worth pursuing. Gradually, I began to realize that what made the process exhilarating was developing increasingly more precise understanding of the gaps in the current literature, iteratively building better models that make slightly more accurate predictions, and most importantly, asking ever more basic questions about the phenomena we observe.

Each time I struggled, a weekly meeting with one of my advisers tended to remind me of the bigger picture. In addition to providing advice on my specific projects, both Ken and Jon

displayed amazing foresight in helping me foster collaborations with people both within and outside the department. Finally, my advisers have been champions for open science, encouraging their lab members to collect larger amounts of data, to replicate their own findings, to share code and data with other labs, and most importantly, to ensure that our interpretations of data stood the test of many sanity checks. It is their passion for getting to the bottom of scientific problems – their deep curiosity about the truth – that gives me the courage to present the flaws, limitations and mysteries in my projects, for those are often the most interesting parts.

Robert Wilson is the first person I worked with at Princeton and perhaps taught me the most about how to build computational models for neuroscience research. Bob guided me in building simple mathematical models of decisions that I could "play" with, taught me how to simulate data, recover the parameters from simulated data, and most importantly, observe by eye how a model behaves as its parameters change, so that I could develop an intuitive understanding of its workings. His kindness, sense of humor and knack for explaining complex ideas have guided me all throughout graduate school.

Michael Shvartsman, as a postdoc in Jon's lab, has played an important role in my development as a cognitive scientist and modeler. It was thanks to his guidance that I had my first experiences of using a mathematical model to test my intuitions about the cognitive processes that give rise to a behavioral pattern. Mike encouraged me to find simple ways to express my intuitions as either computational or statistical models, and showed me just how erroneous these intuitions could be. The experience was one of the most edifying I have had in graduate school. Moreover, his incessant efforts to promote statistical rigor in our department

have been inspiring and have helped me appreciate a wealth of complex tools that I would otherwise not have the courage to approach.

More broadly, I would like to thank the members of Norman and Cohen labs, as well as Yael Niv's labs and Uri Hasson's, especially Janice Chan and Christopher Honey, both for teaching me an incredible amount about neuroscience and for providing a wonderful environment that has felt nourishing and intellectually vibrant. These lab mates never failed to provide feedback about my analyses, interpretations of data and presentation skills. Their insights taught me that the mark of a good scientist is not always the ability to present a coherent story, so much as the ability to present one's hypotheses, methods and reasoning in a manner that makes clear where the mysteries lie.

My friends at Princeton, many of whom are within our department, have kept me happy, providing fascinating conversations, fulfilling trips to nature, quiet game nights and baking nights, and emotional support when I needed it most.

Finally, it is practically impossible to adequately thank my family, and especially my parents, for their tremendous role in developing my curiosity about the world. I feel fortunate to have grown up with parents who are themselves curious, voracious readers, and who taught me that there is nothing more exhilarating than learning how to learn and learning how to think. They have not only shaped me intellectually, but have also provided the emotional safety net – the grounding – which gives me the courage to take risks in the world, knowing that there are things more important than failure.

General Introduction

Retrospective Time Perception and the Contextual Change Hypothesis

The recognition that memory is instrumental to the perception of time dates at least to Aristotle's treatise "On Memory and Reminiscence" (350 BCE), in which he wrote that "only those animals which perceive time remember, and the organ whereby they perceive time is also that whereby they remember." In other words, we can only recognize that some time has elapsed if we are able to discriminate between a past moment and the present one, i.e., if we remember moments other than the present one. It has long been known that memory is not a veridical reflection of our experiences – that it can be selective and fragmented, such that some events are encoded more richly than others, while some are not encoded at all (Craik & Lockhart, 2008; Holland & Kensinger, 2010; Sekeres et al., 2016; Stein, Littlefield, Bransford, & Persampieri, 1984; Yonelinas & Ritchey, 2015). If time perception is generated by the same organ as memory, then it follows that perceived duration should also not vary "linearly" with the duration of our experiences – that some experiences should cause time to dilate, while others should cause it to contract. Despite this age-old intuition, an astonishingly small portion of the neuroscience literature on time perception and temporal processing has investigated how animals and humans can estimate durations longer than a few seconds (see Buhusi & Meck, 2005; Merchant, Harrington, & Meck, 2013; Wittmann, 2013, for a review). That is, in most paradigms used to probe the neural mechanisms of time perception, both the duration of the stimuli being timed and the amount of time between stimulus offset and the subject's

report of stimulus duration, have remained on the scale of seconds (though see Ezzyat & Davachi, 2014; Jacobs, Allen, Nguyen, & Fortin, 2013; Noulhiane, Pouthas, Hasboun, Baulac, & Samson, 2007; Pouthas & Perbal, 2004, which are exceptions that have investigated longer durations and are discussed in more detail in Study 1). This suggests that the influence of long-term and episodic memory on time perception has been left unexplored. Although the definitions of long and short-term memory are debated (Jonides et al., 2008; Nee, Berman, Moore, & Jonides, 2008), it is safe to assume that long-term memory influences duration estimates on the scale of minutes, whereas short-term memory dominates on the sub-second and second timescale. Even among neuroscience studies on the scale of milliseconds to seconds, only a subset has manipulated the information content of the interval to be measured, in order to probe the neural mechanisms that mediate subjective time dilation or expansion (e.g., Dirnberger et al., 2012; van Wassenhove, Wittmann, Craig, & Paulus, 2011; Wittmann, van Wassenhove, Craig, & Paulus, 2010)[1].

At the same time, the fields of experimental psychology and psychophysics have produced a wealth of studies investigating the influence of cognitive, emotional and physiological factors on time perception, carefully manipulating the type of information processed during an interval (see Block & Zakay, 2008; Eagleman et al., 2005; Grondin, 2010; Sucala & David, 2012, for

[1] For instance, Dirnberger et al. (2012) found that the degree of negative affect produced by a picture predicted the degree to which its duration would be overestimated relative to a neutral picture, as well as whether it would be recognized later. They found that BOLD activity in the anterior insula and putamen predicted subsequent memory for aversive pictures whose duration was overestimated.

reviews). In this literature, there is overall consensus that both the duration of the interval being timed (milliseconds to seconds vs. minutes) and the experimental procedure that is used to probe time perception can qualitatively change the relationship between an interval's information content and subjective time. One distinction that has helped explain a large part of the variance is the distinction between prospective and retrospective paradigms. In prospective paradigms, subjects must deliberately measure the duration of an interval as part of the task instructions. In retrospective paradigms, subjects experience an interval (during which they might perform a task) without any awareness that its duration will later have to be reported. They are then surprised by being asked to estimate the duration of some clearly delineated period during the experiment. In both types of paradigms, subjects might be asked to compare the given interval's duration with that of another interval from the experiment, to reproduce its duration with the press of a button or to explicitly report its duration in units of time. In prospective paradigms, subjects might also be asked to press a button once some amount of time has elapsed (e.g., "please press a button after 40 seconds have elapsed.")

The distinction between prospective and retrospective time perception appears to have first been drawn by William James' in *The Principles of Psychology* (1890), in his chapter on "The Perception of Time":

> "*In general, a time filled with varied and interesting experiences seems short in passing, but long as we look back*. *On the other hand, a tract of time empty of experiences seems long in passing, but in retrospect short. A week of travel and sight-seeing may subtend an angle more like three weeks in the memory; and a month of sickness hardly yields more memories than a day. The length in retrospect depends obviously on the multitudinousness of the memories which the time affords. Many objects, events, changes, many subdivisions, immediately widen the view as we look back. Emptiness, monotony, familiarity, make it shrivel up.*"
>
> James, 1890

While the details of this conjecture needed further elaboration, it appears, based on decades of experiments, to have been correct in its essentials.

Factors that influence prospective duration estimates

Experiments employing prospective paradigms have consistently found that parametrically increasing the number of stimuli to be processed during an interval or the difficulty of the task performed on those stimuli results in a corresponding *decrease* in the estimated duration of the interval (Block & Zakay, 2001; Hicks, Miller, & Kinsbourne, 1976; Macar, Grondin, & Casini, 1994; Zakay, 1992, 1998). These findings have been interpreted to mean that measuring duration involves an accumulation process that does not occur automatically, but requires the same attention and working memory resources as the distractor task: the more involving the distractor task becomes, the less time is "integrated" into the representation of the current interval's duration (Zakay & Block, 1997, 2004). In fact, performance on the time estimation task appears to trade off with performance on the distractor task, consistent with the interpretation that they are competing for the same limited cognitive resources (Brown, 1997; Hicks et al., 1976).

Factors that influence retrospective duration estimates

Unlike prospective duration estimates, retrospective duration estimates do not appear to vary with the difficulty of the distractor task performed during the interval. Studies that directly compared the effect of increasing task difficulty on both prospective and retrospective estimates of the exact same interval (between-subject designs) found that task difficulty

impacted prospective estimates but not retrospective ones (Block, 1992; Block & Reed, 1978; Hicks et al., 1976). However, some studies have found that retrospective duration estimates increased with the number and complexity of the stimuli processed during an interval (Block, 1974; Ornstein, 1969; Underwood & Swain, 1973). Such effects are in the opposite direction of those seen for prospective estimates, and originally led to the hypothesis that retrospective duration estimates lengthen with the amount of information that can be remembered from the interval, such that the more memories come to mind from an interval, the longer it will appear in retrospect (Ornstein, 1969).

Despite the intuitive nature of this hypothesis, further experiments have painted a more elaborate picture. Although the number of *stimuli* remembered from an interval does not predict retrospective duration estimates, it appears that the number of larger scale changes in contextual factors does (Block, 1985; Sahakyan & Smith, 2014).

While the term "context" is difficult to define, in long-term memory paradigms it is generally operationalized as any factor other than the stimuli that participants must respond to, but which may nevertheless be encoded in memory. In an experiment where a subject must memorize a list of words, contextual features could include the room in which the experiment takes place, the identity of the experimenter, the subject's current thoughts or emotions, as well as the task that the subject is performing on the stimuli. A property uniting all of these features is that they change more slowly than the stimuli themselves and thus can be associated with multiple stimuli that occur close together in time (Manning, Kahana, & Norman, 2014; Polyn & Kahana, 2008). Moreover, the stimuli themselves could evoke images or memories that are then incorporated into a subject's mental context (Howard & Kahana, 2002).

For instance, while the word “cat” may be a stimulus to be processed, the memories of cats that the stimulus evokes can become part of the subject’s “mental context” and can persist even while the subject is processing the next stimulus, e.g., “pond”.

Studies of retrospective duration have consistently found that changes in contextual features, such as the spatial environment (Block, 1982), the type of task performed on the stimuli (Block, 1992; Block & Reed, 1978; Sahakyan & Smith, 2014), emotional state (Pollatos, Laubrock, & Wittmann, 2014), as well as unexpected changes that segment the interval into multiple parts (Block, 1978; Faber & Gennari, 2015; Poynter, 1983; Zakay, Tsal, Moses, & Shahar, 1994) cause the interval duration to be overestimated. A consistent finding across all the above studies was that manipulations that induce contextual change did not improve recognition memory or free recall for the stimuli presented during the interval, arguing against the interpretation that large contextual changes simply engage more attention, promoting deeper processing and better memory for the whole interval. In fact, both Block (1992) and Block & Reed (1978) showed that instructing subjects to perform tasks that promoted “deeper” stimulus processing and resulted in better subsequent memory did not affect duration estimates, whereas changing the type of task performed during an interval (promoting segmentation of the interval) lengthened retrospective estimates. Sahakyan & Smith (2014) confirmed these results, showing that retrospective duration estimates were not correlated with the number of stimuli remembered from an interval in any of their experiments, whereas task instructions that required changes in the type of processing consistently lengthened duration estimates. Importantly, although contextual changes were not accompanied by better recall or recognition of the stimuli from the interval, they did result in greater separation of

stimulus memories by context, such that subjects were less likely to confuse which stimuli occurred before vs. after the contextual change (Block, 1982; Sahakyan & Smith, 2014).

To summarize, it has been shown that retrospective duration estimates do not increase monotonically with the number of stimuli remembered from an interval, but rather, may be a function of the degree of chunking (segmentation) of the memories from the interval. Such chunking may scale with the number and degree of contextual changes during an interval. Faber & Gennari (2015) provided evidence for this idea by having an independent group of participants rate the number of segments they perceived in dynamic visual stimuli (geometric shapes which either moved in predictable or increasingly unpredictable ways throughout the video), as well as the similarity of the segments to each other. They found that retrospective duration estimates for the videos correlated not only with the degree of segmentation (the number of "boundaries") in each video, but also with the degree of dissimilarity between the segments.

Note that prospective paradigms do not show the opposite effects from retrospective ones: it is not the case that more engaging stimuli cause shorter time estimates prospectively, but longer estimates retrospectively. Rather, it appears that prospective and retrospective estimates are influenced by different experimental factors. Prospective estimates appear strongly correlated with how much attention is devoted to the task of estimating time. Retrospective estimates appear to be driven by change in contextual factors during the interval.

Novelty of Study 1 with Respect to Existing Literature

As mentioned above, almost the entirety of the neuroscience literature on time perception has focused on prospective paradigms on the scale of seconds, leaving the neural mechanisms of retrospective timing entirely unexplored. In Study 1, we sought to bridge the cognitive neuroscience literature with the experimental psychology literature by investigating the neural substrates that mediate retrospective time perception on the scale of minutes. We used functional magnetic resonance imaging (fMRI) to track changes in neural activity while subjects listened to a 25-minute radio story, and related this neural activity at encoding to subsequent duration judgments.

With this study, we sought to fill several gaps in the neuroscience literature: 1) a study in which the degree and number of contextual changes is varied across intervals in a naturalistic way, and in which the effects of contextual change on retrospective duration perception are assessed both within and across subjects, 2) a study in which subjects are not aware that duration must be measured, and which probes intervals on the scale of minutes, such that long-term memory is necessary for estimating duration.

Several obstacles had rendered the design of such a study challenging in the past. First, even in the experimental psychology literature, most retrospective studies only obtained a single duration estimate per subject. This is because, after the first duration estimate, the subject becomes aware of the necessity to estimate durations and the paradigm becomes prospective. Faber & Gennari (2015) circumvented this issue by presenting subjects with multiple videos (on which they had to perform some "cover" task that was not of experimental

interest), and only asking them to judge the durations of the videos at the end of the experiment. However, this was only possible because each video was very short in duration (maximum 14 seconds). Asking participants to estimate the durations of multiple intervals several minutes in length would not be practical in a laboratory experiment. To exacerbate matters, retrospective experiments are practically impossible to conduct in animals. Animal subjects require many trials of training to learn that the duration of an interval must be indicated in order to earn a reward, such that the animal by definition learns that the temporal dimension of stimuli must be attended to, making the experiment prospective.

To address these issues, we developed an experimental design that enables us to obtain dozens of retrospective duration estimates within the same subject. Human subjects listened to a 25-minute-long radio drama while in the fMRI scanner. After coming out of the scanner, they were asked to estimate how much time passed between pairs of clips from the story. We used the clips to delineate 24 two-minute intervals and 19 six-minute intervals in the story. These long intervals overlapped with each other, enabling us to probe 43 minutes-long intervals within the time frame of a single laboratory experiment. Moreover, because the auditory story is a far more engaging than laboratory-designed images or words, subjects were able to recall when in the story each clip had occurred with remarkable accuracy. This enabled them to remember which interval in the story the two clips delineated, without the experimenter introducing any artificial markers of when the interval began and ended. Thus, our paradigm enabled us to collect multiple duration judgments without affecting the perception of the story during encoding (without raising attention to time).

A second obstacle that has precluded neuroimaging studies from investigating intervals on the order of minutes is the presence of slowly varying sources of physiological and scanner noise, which are confounded with neural activity that changes on the scale of minutes. To counteract this problem, we developed a preprocessing pipeline that strives to optimize the trade-off between preserving neural signal that changes slowly over time and filtering out the greatest amount of low-frequency noise. For this purpose, we leveraged the finding from Uri Hasson's lab that when people listen to stories or watch films, their BOLD signal time courses become highly synchronized across subjects in brain regions involved in processing the story (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004). However, when low-frequency scanner noise is not filtered out, spurious correlations across subjects can also arise in regions that contain cerebrospinal fluid (CSF) and therefore cannot be caused by neural activity. We used synchronization of the signal across subjects in regions of interest (known to be involved in story processing) as a measure of signal quality, and synchronization in regions of no interest (such as CSF and white matter) as a measure of noise. Based on these two metrics, we picked the high-pass filter that removed the greatest amount of noise, while still enabling us to measure neural activity changing on the time scale of minutes.

Hypothesis in Study 1

Based on the literature reviewed above, we hypothesized that in order to estimate how much time elapsed between two clips from the story, subjects would retrieve memories associated with the two clips, and assess the degree of dissimilarity between them. We hypothesized that contextual changes would be the main factor influencing the perceived

dissimilarity between the two memories, such that subjects' duration estimates would be proportional to the degree of contextual change between the two clips. A plethora of studies reviewed in Study 1 have found that the medial temporal lobes (MTL) and the lateral prefrontal cortex (PFC) play an important role in encoding the context (spatial, social or temporal) in which stimuli are embedded. The MTL appears particularly important for forming long-term memories of which stimuli occurred in which context, while the PFC appears particularly important for using contextual information to guide behavior (although such statements are certainly oversimplified). We therefore used the pattern of neural activity in these regions (measured by fMRI) at the time when a clip was first played in the story as a proxy for the subject's encoding of contextual information during that moment in the story. We expected that, if subjects use the contextual change between two story clips to judge how much time passed between them, then change in neural activity patterns (in regions representing context) might correlate with retrospective duration estimates.

Main Findings of Study 1

Several findings in Study 1 lend support to the above hypotheses. First, we asked a separate group of participants to press a button whenever they perceived an "event boundary" in the story (whenever one event had just ended and a new event was beginning). Subjects tended to perceive boundaries whenever several factors, such as the location, the time of day, the characters present and the topic of conversation changed in the story. We found that the number of event boundaries during an interval was highly correlated with how long that interval was perceived to be in retrospect, despite all intervals having the same objective

duration (either two minutes or six minutes). This replicated previous findings that contextual changes caused duration estimates to lengthen in retrospect. Second, using both within-subject and across-subject analyses, we found that neural pattern dissimilarity in the right anterior temporal lobe, particularly the right entorhinal cortex, was correlated with subjects' retrospective duration estimates. In other words, the more different patterns in these regions appeared when the two clips were first played in the story, the more time subjects believed passed between the two clips (despite all clips being two minutes apart). We found these effects using both region-of-interest and whole-brain analyses, demonstrating that the correlations between brain and behavior really were strongest in the right anterior temporal lobe, even when all possible regions were considered. Third, we sought to rule out the concern that our findings were driven by perceptual similarity between the pairs of story clips, rather than similarity in abstract contextual factors. The concern was that subjects might be basing their duration estimates on how different the two clips sound during the memory test (e.g., which characters are speaking, which background music is playing), which does not require them to remember the narrative context associated with each clip. If patterns of activity in entorhinal cortex reflect any of the perceptual content of the clip, then pattern change in entorhinal cortex might correlate with duration estimates because both are influenced by perceptual, rather than contextual, similarity. To rule out this concern, we asked a separate group of "naïve" subjects who had never heard the story to estimate how much time passed between each pair of clips, without any knowledge of what happened in the story between the two clips. We found that naïve estimates of duration were highly reliable across subjects, suggesting that they used a similar strategy to estimate how much time passed between two

clips (as we expected, naïve subjects reported basing their duration estimates on how similar the two clips sounded, such as which music was playing, which characters were speaking, the topic discussed in the two clips, etc.). However, including these naïve estimates in the model did not explain the correlation between duration estimates and neural pattern dissimilarity. Moreover, naïve estimates could not account for the correlation between the number of event boundaries and duration estimates. We did find a significant correlation between naïve duration estimates and the duration estimates of our original subjects (who had heard the story), but we also found that there a significant component of our original subjects' behavior that the naïve estimates could not explain. In other words, as far as we could ascertain, the relationship between neural pattern change in entorhinal cortex and retrospective duration estimates did not appear to be caused by the perceptual similarity between the two clips delimiting each interval. Rather, the relationship seemed to be impacted by subjects' memory for what had happened between the two clips in the story.

Relationship between Study 1 and Study 2

In Study 1, we found a strong correlation between the number of "event boundaries" during an interval and retrospective duration estimates for that interval. This finding is in line with all the studies reviewed above, which had shown that changes that segment an interval into multiple parts lengthen retrospective duration judgments. Above, we mentioned that these same studies had also found that greater contextual changes led to greater separation of memories by context (i.e., reduction of interference), such that subjects successfully remembered whether a stimulus occurred before or after the change. Such findings suggest

that stimuli that occurred in different contexts might be coded more dissimilarly in long-term memory, which results in the "illusion" that they happened further apart than they actually did (in Study 1, subjects consistently overestimated the durations of 2-minute intervals that contained more than one event boundary). Since we only collected fMRI data during the encoding of the story, we cannot claim that pairs of clips that elicited similar patterns in the anterior temporal lobe at encoding are also the same pairs that will appear similar at retrieval, during the time perception test. However, our finding that similarity at encoding correlates with duration estimates at retrieval (which takes place about 30 minutes later) is at least suggestive that moments in the story that are represented similarly in these regions during encoding will also appear similar at retrieval. Future studies will be needed to track how the representations of the clips change between encoding and retrieval, whether separation by context is accentuated over time, and whether the regions that predict duration estimates at encoding are the same as those that predict them at retrieval.

Previous studies that specifically examined how event boundaries influence memories for complex naturalistic stimuli (such as stories and movies) have found that information presented before a boundary becomes less accessible after the boundary in the short term (Morrow, Bower, & Greenspan, 1989; Rinck & Bower, 2000; Speer & Zacks, 2005; Swallow, Zacks, & Abrams, 2009; Zwaan, 1996), and that memories appear to cluster within events in the long term (DuBrow & Davachi, 2013; Ezzyat & Davachi, 2011). For instance, Ezzyat & Davachi (2011) found that inserting boundary keywords (such as "a while later") into a narrative tended to weaken the association between sentences across the boundary in long-term memory compared with control keywords ("a moment later"). This suggests that event boundaries in

naturalistic stimuli play the same role as contextual changes in more artificial paradigms, weakening the associations between items across boundaries, or strengthening the associations between items presented during the same "event" (within the same context).

Together with existing literature, our findings in Study 1 point to "mental context" as a strong organizing principle for long-term memory and suggest that time estimates might reflect the organization of memories by context.

In Study 1, the contextual factors that varied across event boundaries were similar to those that had been varied in previous studies: location, time of day, the characters present in a scene, the topic of conversation, etc. In Study 2, we were interested in how organisms learn to detect boundaries in the first place. In other words, how does repeated experience with a certain type of change influence whether that change is viewed as an important boundary?

One idea from *event segmentation theory* (Zacks, Speer, Swallow, Braver, & Reynolds, 2007) is that boundaries arise when we perceive stimuli or perform actions that are less predictable based on the stimuli or actions that occurred just before. For instance, actions associated with the task of "making coffee" (reaching for a mug, grinding coffee beans, retrieving milk from the fridge) often occur close together in time, such that after some experience watching people making coffee, an observer can predict which action will occur next based on the previous two. On the other hand, the last action of preparing the coffee (pouring it into the mug) might be followed equally often by reaching for yogurt in the fridge or by reaching for the newspaper. Thus, the transition between stimuli or actions within events might occur more frequently than transitions across events. Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick (2013) shed further light on this question by showing that human subjects learn to detect boundaries even

when they experience transitions that had the same probability of occurring as transitions within events, so long as the transition occurs between two stimuli that are associated with different clusters of stimuli. In their experiment, subjects performed a cover task while viewing a series of fractals. Unbeknownst to the subjects, there were three "communities" of fractals, such that most of the fractals within a community were only connected with each other, but two of the fractals had a connection to the other communities. In other words, subjects experienced far more transitions between fractals in the same community. However, when they experienced transitions between communities, those transitions were as likely as transitions within the cluster, given the fractal on the previous trial. This study demonstrated that a key feature of event boundaries is that they constitute transitions between stimuli that have different temporal or semantic associations, but that the "boundary stimulus" need not necessarily be surprising given the stimulus that occurred just before. Moreover, using fMRI to track the neural representation of each fractal, the authors found that the representations of fractals in the same communities grew more similar to each other over the course of statistical learning (in the left anterior temporal lobe, super temporal gyrus, insula and inferior frontal gyrus). This suggests that stimuli with similar temporal associations become similarly represented, despite any differences in perceptual similarity within and across clusters at the start of the experiment.

The above findings suggest that contextual boundaries may arise when we experience stimuli that have relatively different associations from the stimuli that occurred previously. In Study 2, we investigated the principles that govern how humans learn different associations with different contexts, in paradigms where the context not only predicts upcoming stimuli, but

also which actions are appropriate. As mentioned above, previous studies have shown that changes in the task performed on the exact same stimuli lengthened duration estimates (Block, 1992; Block & Reed, 1978), suggesting that contextual boundaries occur not only in response to changes in stimulus associations, but also in response to changes in the actions that are appropriate to those stimuli (stimulus-response mappings). Specifically, we were interested in how similarly two contexts would be perceived as a function of both the stimuli and actions that they predict. We tested whether principles of Bayesian inference, which may explain how people cluster their observations by context in perceptual learning (Gershman & Niv, 2013), could also explain how people cluster their observations in the domain of task learning (where a contextual cue can predicts stimuli, responses and stimulus-response mappings). According to these principles, category learning should obey Occam's Razor, preferring simpler explanations of data to more complex ones. In the case of grouping vs. separating memories by context, this suggests that observations should be grouped across contexts as much as possible, resulting in general predictions that apply across various contexts, until enough evidence is accumulated that two contexts make truly distinct predictions about upcoming stimuli and actions. To test this idea, we parametrically manipulated the degree to which two contexts made similar predictions, aiming to verify whether expectations would become separated by context only when they made sufficiently different predictions.

Novelty of Study 2 with Respect to Existing Literature

In the Discussion of Study 2, we review two branches of the literature, both of which suggest that principles of Bayesian inference might explain how predictions are combined

across contextual cues. First, studies of perceptual cue combination have tested how sensory cues (such as binocular disparity, object motion, occlusion) are combined to determine a property of an object, such as its depth in a visual scene. Such studies have found that, when the various cues provide slightly different information, organisms average depth information across cues, weighting each cue by its reliability (Ernst & Banks, 2002; Gepshtein and Banks, 2003). There is also evidence that people assign increasingly less weight to a cue as its conflict with the other cues increases (Knill, 2007; Girshick and Banks, 2009). These results are consistent with the hypothesis that cue combination amounts to a causal inference problem: information from a specific cue is averaged with the other cues to the degree that it seems generated by the same source, but discarded to the degree that it appears to be an outlier or to stem from a

different source (Landy et al., 1995; Hospedales and Vijayakumar, 2009).

The second line of evidence comes from the literature on category learning, showing a bias towards learning simpler and smaller numbers of concepts, unless a more complex representation is crucial for explaining the observations (Anderson, 1991; Clapper and Bower, 1994; Pothos and Chater, 2002; Love et al., 2004; Sanborn et al., 2010; Goldstone, 1994; Sloutsky, 2003). Gershman and Niv (2013) tested whether the same simplicity principle would govern how one stimulus property (the color of circles on a screen) would influence people's estimates of another stimulus property (the number of circles), even though subjects were only asked to indicate the number of circles on each trial (the instructions said nothing about color). They found that when both colors predicted similar distributions of the number of circles, subjects' estimates tended to be biased towards the mean of the two distributions. On the

other hand, when the two colors predicted less overlapping distributions, subjects' estimates were significantly less biased towards the mean, suggesting that the similarity of predictions between the two colors determined the degree to which their predictions were combined, just as predicted by the literature on perceptual cue combination.

The first branch of the literature provides evidence that Bayesian principles govern how information is combined across contextual cues to disambiguate stimuli within single trials, while the second branch suggests that they also govern how people learn, across many trials, to use contextual cues to categorize stimuli (it is possible that, after category learning has taken place, changes in these cues may be interpreted as event boundaries, though such hypotheses need to be tested experimentally).

An unexplored question in both of the above literatures is whether probability distributions predicted by two cues can be "merged" even when differentiating the cues is crucial to behaving appropriately. In the above experiments, the cues whose predictions were combined to inform a perceptual decision (the number of circles on screen or the depth of an object in a visual scene) did not influence the way in which the decision was reached. In fact, once the cues were integrated into the perceptual representation, information about the individual cues could be discarded without any loss in performance (e.g., if the two colors predicted very similar distributions of circle numbers, then the color dimension could effectively be ignored). However, in real-world situations, contextual features (such as our spatial or social environment) often determine not only the probabilities of the stimuli we might see next, but also change which actions are appropriate to those stimuli. In Study 2, we used a context-dependent decision making task (see Study 2 for design details), in which two cues, A and B, not

only predicted which stimulus is more likely to appear next (X or Y) and which response is more likely (left or right), but also deterministically changed the correct response to the stimulus. We then tested whether stimulus or response probabilities could be merged across cues over the course of many trials, even when the response rules signaled by the two cues must remain distinct in order to guide behavior.

Findings and Contributions of Study 2

The results of Study 2 provide strong evidence that "weak fusion" of cue predictions can occur in tasks in which distinguishing the cues is necessary for correct responding. ("Weak fusion" is a term employed in the literature on cue combination to signify that individual cue information is retained even once the cue has been averaged with other cues to form a unified percept; Landy et al., 1995). In our experiments, we parametrically manipulated the trial frequencies across subjects, such that the two cues predicted similar response probabilities (but opposite stimulus probabilities) for some subjects and similar stimulus probabilities (but opposite response probabilities) for other subjects. To analyze how the cue affected people's predictions about the stimulus and response, we fit a drift diffusion model we had built to subjects' behavior. The model's starting point (or Bayesian "prior") parameter effectively captured the degree to which subjects were faster and more accurate on left vs. right responses after each cue. We found that subjects' average response priors on A cues were consistently biased towards the average response priors on B cues, and that this "merging" of response priors across increased with the similarity of their predictions about the response. These results are in agreement with those of Gershman and Niv (2013). Importantly, merging of

response probabilities occurred despite the fact that subjects kept their representations of the stimulus-response mappings entirely separated by cue. Fitting the drift diffusion model showed that the cue noise parameter, which represents the degree to which behavioral rules are confused across cues, was equally low (if not lower) in the conditions were merging of response probabilities was most apparent, suggesting that response probability information could be combined across cues without any loss in performance and without any merging of behavioral rules.

Besides this finding, our study provides insight into possible mechanisms for merging of probabilities across cues, and suggests that cue-dependent probabilities are combined by changing representations over the course of many trials, rather than combined within individual trials. These insights stem from comparing subjects' reaction time and accuracy behavior on the task with their explicit reports of trial frequencies at debriefing. In addition to measuring subjects' task performance over the course of learning, we asked them to explicitly report the probability of each cue-stimulus combination (AX, AY, BX, BY) at the end of the experiment. To our surprise, subjects' explicit probability estimates also showed merging of response probabilities across cues, such that people's estimates of AX/AY frequency were biased towards the true frequencies of BY/BX trials. In fact, the pattern of merging across these two distinct behavioral measures was so similar that feeding subjects' estimates of trial frequencies into the drift diffusion model, and deriving the optimal starting points for each subject given their estimates and other parameters, showed that subjects' priors were indistinguishable from optimal given their estimates of trial frequencies, at least for 300 online (Mechanical Turk) subjects. The match between subjects' behavior and their probability

estimates argues against the possibility that merging across cues occurs on the level of individual trials. Such mixing of cue probabilities within trials could occur, for instance, due to limitations in working memory capacity during the delay between cue and stimulus, resulting in confusion about which cue to generate predictions from. Thus, it could have been the case that subjects maintained distinct probability estimates for the two cues and only combined the probabilities on individual trials. However, if this were the case, their explicit estimates of probabilities would show separation of probabilities by cue, while their performance on the task would show merging. The fact that merging is visible in explicit estimates, and that behavior appears optimal given these estimates, suggests a different account: the distortions arise from subjects' memory of the probabilities, but given these merged memories, the probabilities appear to be optimally implemented in their decisions on individual trials. To summarize, merging appears to occur at the representational level, rather than during the implementation of separated representations.

Another contribution of Study 2 is that it reveals how differences in cue-stimulus frequencies cause subjects to learn different ways of structuring their observations by context, and that this structure learning transfers even when the frequencies have changed. Examining how the response priors for the two cues evolved over blocks within subjects showed that they were positively correlated with each other when the two cues predicted similar responses ("merge" condition), but negatively correlated when the two cues predicted opposite response and similar stimuli ("split" condition). Analyzing the explicit reports revealed that this negative correlation appeared to be caused by grouping of stimulus, rather than response, probabilities in the split condition. In other words, it appeared that subjects flexibly switched the dimension

(stimulus or response) along which probabilities were merged across cues, depending on which one was most similar between them. To verify whether these correlations were only caused by the task environment or whether they reflected grouping of the observations internally, we ran a follow-up experiment in which subjects were exposed to 200 trials of either the "merge" condition (both cues predict similar responses, opposite stimulus) or the "split" condition (both cues predict similar stimuli, but opposite responses). In the middle of the experiment, the frequencies of BX/BY trials changed to 25-25, such that B no longer predicted either the left or right response (and no longer predicted the X or Y stimulus), whereas the A frequencies remained the same. As a result, in the second part of the experiment, both the "merge" and the "split" subjects saw exactly the same trial frequencies, such that any difference in their behavior must arise from the strategy they learned earlier in the experiment. We found that "merge" subjects continued to show positive correlations between their A and B response priors, while "split" subjects continued to show negative correlations, even though the new trial frequencies did not promote any correlations between A and B frequencies. These results suggest that the correlations between the priors across blocks reflected subjects' learning of how to group observations by cue, and that subjects continued to apply their grouping strategy even when the true frequencies had changed and no longer favored any grouping.

In this way, Study 2 furthers our understanding of how structure is learned in the domain of task representations. It shows that subjects can group either stimulus or response probabilities across cues, depending on which dimension affords better performance, and that this grouping does not come at the expense of maintaining separate values of the probabilities for each cue, since subjects were able to adapt the mean value of their response prior for B cues in response

to the change in B frequencies, even though they continued to group the priors in the same way they had learned before. All of these effects are consistent with the idea that humans perform "weak fusion", combining information across cues to improve the reliability of predictions, while still maintaining individual cue information when it is needed for correct performance.

Finally, Study 2 also contributes a new angle to the literature on context-dependent decision-making. The ability to adaptively change behavior as the situation changes, and to perform actions that are appropriate for a given context, has been shown to be remarkably fragile, and impaired in a variety of patient populations with deficits of "executive function" (Barch et al., 1998; Braver & Cohen, 2000; Cohen, Barch, Carter, & Servan-Schreiber, 1999; Schwartz, Reed, Montgomery, Palmer, & Mayer, 1991; Servan-Schreiber, Cohen, & Steingard, 1996; Sirigu et al., 1996; Zalla, Verlut, Franck, Puzenat, & Sirigu, 2004). Disorders of executive function, in which context does not sufficiently influence behavior, have been investigated with tasks in which subjects should have already learned that different behaviors are appropriate for different contexts, but fail to apply this knowledge, possibly due to deficits in maintaining or using contextual information. In Study 2, we turn instead to the question of which algorithms people use to learn whether context-specific representations in the first place. We show that even when subjects optimally implement different behavioral rules for each context, they can still merge predictions across contexts, possibly in order to improve the robustness of predictions in the face of limited data. Accounting for merging predictions across contexts becomes important if a researcher wishes to distinguish whether deficits in context-

dependence arise because subjects never learned sufficiently different representations for each context, or whether subjects did learn perfectly separate representations but are unable to implement them correctly, due to failure in context maintenance, for example.

This distinction becomes crucial in Study 3, where we demonstrate that behavioral patterns that have been taken as signature of context processing failure can also arise from distortions in probability learning.

Relationship between Study 2 and Study 3

In Study 2, we investigated how subjects learn to use contextual cues to predict future stimuli and actions, and how the similarity of cue predictions influences the degree to which subjects form cue-dependent priors. In Study 3, we examined the possibility that the associations between contextual cues and stimuli could be bidirectional, such that stimuli could also be informative about which cue is more likely, especially in the face of uncertainty about the cue. For instance, a subject who forgot whether they saw an A or B cue on a particular trial, but who knows that X is more likely after A whereas Y is more likely after B, could still infer that A is more likely than B if an X stimulus appears.

A wealth of studies has hypothesized that context-dependent behavior could arise through two different strategies: a proactive strategy, in which the cue is processed in advance and used to prepare the decision on the stimulus, and a reactive strategy, in which the cue might be noisily represented, but in which the stimulus can retroactively help reduce uncertainty about the cue. This hypothesis has been termed the “dual mechanisms of control” framework (Braver, 2012; Braver et al., 2007). In Study 3, we sought to test the tenets of this framework,

investigating whether cue predictions are always used in advance of the stimulus, or whether there is strategic variability, such that the stimulus might also be used to constrain the probability of the cue.

To do this, we used the same drift diffusion model as in Study 2, but added a "reactive" version of the model: rather than using the cue to set the prior for the decision on the probe, the reactive model uses the probe to set the starting point for retrieving the memory of the cue. We then used simulations to characterize which types of behavioral patterns the two models could generate. Specifically, we explored how changing the model parameters influenced the model's performance on AY relative to BX trials, since the relative error rates and correct reaction times on AY and BX trials have been used as a behavioral index of proactive and reactive control.

Within the purely proactive model, we found that two sets of parameters could affect the behavioral index: the amount of cue noise (uncertainty about the cue) relative to the drift rate (certainty about the stimulus), and the degree of similarity between A and B starting points. So long as cue noise was above 0, we found that increasing the similarity between A and B starting points led the model to generate progressively more BX errors than AY errors, which has been taken as a signature of reactive strategy. However, in Study 2, we had found that similarity between A and B starting points could result from merging of probabilities across cues over the course of many trials. Thus, our simulations show that behavioral variability that has been interpreted as variability in strategy within trials could in fact arise from variability in learning predictions across trials.

In the General Discussion section, we review additional findings from Study 3 that provide a potential solution to this problem, pointing to reaction times as a more reliable indicator of within-trial dynamics.

References

Barch, D. M., Cohen, J. D., Servan-Schreiber, D., Braver, T. S., Stratta, P., Daneluzzo, E., … Rossi, A. (1998). Schizophrenic deficits in the processing of context. *Archives of General Psychiatry*, *55*(2), 187–188. doi:10.1001/archpsyc.55.2.186

Block, R. A. (1974). Memory and the experience of duration in retrospect. *Memory & Cognition*, *2*(1), 153–160. doi:10.3758/BF03197508

Block, R. A. (1978). Remembered duration: Effects of event and sequence complexity. *Memory & Cognition*. doi:10.3758/BF03197462

Block, R. A. (1982). Temporal judgments and contextual change. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *8*(6), 530–544. doi:10.1037/0278-7393.8.6.530

Block, R. A. (1985). Contextual coding in memory: Studies of remembered duration. In *Time, Mind and Behavior* (pp. 169–178).

Block, R. A. (1992). Prospective and retrospective duration judgment: The role of information processing and memory. In F. Macar, V. Pouthas, & W. J. Friedman (Eds.), *Time, actions and cognition: Towards bridging the gap* (pp. 141–152). Dordrecht, The Netherlands: Kluwer Academic.

Block, R. A., & Reed, M. A. (1978). Remembered duration: Evidence for a contextual-change hypothesis. *Journal of Experimental Psychology: Human Learning & Memory*. doi:10.1037/0278-7393.4.6.656

Block, R. A., & Zakay, D. (2001). Retrospective and prospective timing: Memory, attention and consciousness. In C. Hoerl & T. McCormark (Eds.), *Time and Memory: Issues in Philosophy and Psychology* (pp. 59–76). Oxford, England: Oxford University Press. Retrieved from http://www.montana.edu/rblock/documents/papers/BlockZakay2001a.pdf

Block, R. A., & Zakay, D. (2008). Timing and remembering the past, the present, and the future. *Psychology of Time*. doi:10.1016/B978-0-08046-977-5.00012-0

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. doi:10.1016/j.tics.2011.12.010

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In *Control of cognitive processes Attention and performance XVIII* (Vol. XVIII, pp. 713–737). doi:10.1016/S0165-0173(03)00143-7

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. *Variation in Working Memory*. doi:10.3233/978-1-60750-044-5-250

Brown, S. W. (1997). Attentional resources in timing: Interference effects in concurrent temporal and nontemporal working memory tasks. *Perception & Psychophysics*, *59*(7), 1118–1140. doi:10.3758/BF03205526

Buhusi, C. V, & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews. Neuroscience*, *6*(10), 755–65. doi:10.1038/nrn1764

Cohen, J. D., Barch, D. M., Carter, C., & Servan-Schreiber, D. (1999). Context-processing deficits in schizophrenia: converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology*, *108*(1), 120–133. doi:10.1037/0021-843X.108.1.120

Craik, F. I. M., & Lockhart, R. S. (2008). Levels of Processing and Zinchenko's Approach to Memory Research. *Journal of Russian and East European Psychology*, *46*(6), 52–60. doi:10.2753/RPO1061-0405460605

Dirnberger, G., Hesselmann, G., Roiser, J. P., Preminger, S., Jahanshahi, M., & Paz, R. (2012). Give it time: Neural evidence for distorted time perception and enhanced memory encoding in emotional situations. *NeuroImage*, *63*(1), 591–599. doi:10.1016/j.neuroimage.2012.06.041

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, *142*(4), 1277–86. doi:10.1037/a0034024

Eagleman, D. M., Tse, P. U., Buonomano, D., Janssen, P., Nobre, A. C., & Holcombe, A. O. (2005). Time and the brain: how subjective time relates to neural time. *J Neurosci*, *25*(45), 10369–10371. doi:10.1523/JNEUROSCI.3487-05.2005

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science : A Journal of the American Psychological Society / APS*, *22*, 243–252. doi:10.1177/0956797610393742

Ezzyat, Y., & Davachi, L. (2014). Similarity breeds proximity: Pattern similarity within and across

contexts is related to later mnemonic judgments of temporal proximity. *Neuron*, *81*, 1179–1189. doi:10.1016/j.neuron.2014.01.042

Faber, M., & Gennari, S. P. (2015). In search of lost time: Reconstructing the unfolding of events from memory. *Cognition*, *143*, 193–202. doi:10.1016/j.cognition.2015.06.014

Gershman, S. J., & Niv, Y. (2013). Perceptual estimation obeys Occam's razor. *Frontiers in Psychology*, *4*(SEP). doi:10.3389/fpsyg.2013.00623

Goldstone, R. L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition*, *52*(2), 125–157. doi:10.1016/0010-0277(94)90065-5

Grondin, S. (2010). Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception & Psychophysics*, *72*(3), 561–582. doi:10.3758/APP.72.3.561

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science (New York, N.Y.)*, *303*(5664), 1634–1640. doi:10.1126/science.1089506

Hicks, R. E., Miller, G. W., & Kinsbourne, M. (1976). Prospective and retrospective judgments of time as a function of amount of information processed. *The American Journal of Psychology*, *89*(4), 719–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1020767

Holland, A. C., & Kensinger, E. A. (2010). Emotion and autobiographical memory. *Physics of Life Reviews*. doi:10.1016/j.plrev.2010.01.006

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. doi:10.1006/jmps.2001.1388

Jacobs, N. S., Allen, T. A., Nguyen, N., & Fortin, N. J. (2013). Critical role of the hippocampus in memory for elapsed time. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(34), 13888–93. doi:10.1523/JNEUROSCI.1733-13.2013

James, W. (1890). *The principles of psychology (Vols. 1 & 2)*. *New York Holt* (Vol. 118). doi:10.1037/10538-000

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The Mind and Brain of Short-Term Memory. *Annual Review of Psychology*, *59*(1), 193–224. doi:10.1146/annurev.psych.59.103006.093615

Macar, F., Grondin, S., & Casini, L. (1994). Controlled attention sharing influences time estimation. *Memory & Cognition*, *22*(6), 673–686. doi:10.3758/BF03209252

Manning, J. R., Kahana, M. J., & Norman, K. A. (2014). The role of context in episodic memory. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The Cognitive Neurosciences* (V., pp. 557–566).

Cambridge, MA: MIT Press.

Merchant, H., Harrington, D. L., & Meck, W. H. (2013). Neural Basis of the Perception and Estimation of Time. *Annual Review Neuroscience*, *36*, 313–36. doi:10.1146/annurev-neuro-062012-170349

Morrow, D. G., Bower, G. H., & Greenspan, S. L. (1989). Updating situation models during narrative comprehension. *Journal of Memory and Language*, *28*(3), 292–312. doi:10.1016/0749-596X(89)90035-1

Nee, D. E., Berman, M. G., Moore, K. S., & Jonides, J. (2008). Neuroscientific evidence about the distinction between short- and long-term memory. *Current Directions in Psychological Science*, *17*(2), 102–106. doi:10.1111/j.1467-8721.2008.00557.x

Noulhiane, M., Pouthas, V., Hasboun, D., Baulac, M., & Samson, S. (2007). Role of the medial temporal lobe in time estimation in the range of minutes. *Neuroreport*, *18*(10), 1035–8. doi:10.1097/WNR.0b013e3281668be1

Ornstein, R. E. (1969). On the experience of time. *Hamondsworth (UK): Penguin Books*.

Pollatos, O., Laubrock, J., & Wittmann, M. (2014). Interoceptive focus shapes the experience of time. *PLoS ONE*, *9*(1). doi:10.1371/journal.pone.0086934

Polyn, S. M., & Kahana, M. J. (2008). Memory search and the neural representation of context. *Trends in Cognitive Sciences*, *12*(1), 24–30. doi:10.1016/j.tics.2007.10.010

Pouthas, V., & Perbal, S. (2004). Time perception depends on accurate clock mechanisms as well as unimpaired attention and memory processes. In *Acta Neurobiologiae Experimentalis* (Vol. 64, pp. 367–385).

Poynter, W. D. (1983). Duration judgment and the segmentation of experience. *Memory & Cognition*, *11*, 77– 82.

Rinck, M., & Bower, G. H. (2000). Temporal and spatial distance in situation models. *Memory & Cognition*, *28*(8), 1310–1320. doi:10.3758/BF03211832

Sahakyan, L., & Smith, J. R. (2014). “A long time ago, in a context far, far away”: retrospective time estimates and internal context change. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(1), 86–93. doi:10.1037/a0034250

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Publishing Group*. doi:10.1038/nn.3331

Schwartz, M. F., Reed, E. S., Montgomery, M. W., Palmer, C., & Mayer, N. H. (1991). *The quantitative description of Action Disorganisation after brain damage: A case study*. *Cognitive Neuropsychology* (Vol. 8). doi:10.1080/02643299108253379

Sekeres, M. J., Bonasia, K., St-Laurent, M., Pishdadian, S., Winocur, G., Grady, C., & Moscovitch, M. (2016). Recovering and preventing loss of detailed memory: differential rates of forgetting for detail types in episodic memory. *Learning & Memory*, *23*(2), 72–82. doi:10.1101/lm.039057.115

Servan-Schreiber, D., Cohen, J. D., & Steingard, S. (1996). Schizophrenic deficits in the processing of context: A test of a theoretical model. *Archives of General Psychiatry*, *53*, 1105–1112.

Sirigu, A., Zalla, T., Pillon, B., Grafman, J., Agid, Y., & Dubois, B. (1996). Encoding of sequence and boundaries of scripts following prefrontal lesions. *Cortex*, *32*(2), 297–310. doi:10.1016/S0010-9452(96)80052-9

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*. doi:10.1016/S1364-6613(03)00109-8

Speer, N. K., & Zacks, J. M. (2005). Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of Memory and Language*, *53*(1), 125–140. doi:10.1016/j.jml.2005.02.009

Stein, B. S., Littlefield, J., Bransford, J. D., & Persampieri, M. (1984). Elaboration and knowledge acquisition. *Memory & Cognition*, *12*(5), 522–9. doi:10.3758/BF03198315

Sucala, M., & David, D. (2012). Slowing Down the Clock: A Review of Experimental Studies Investigating Psychological Time Dilation. *The Journal of General Psychology*, *139*(4), 230–243. doi:10.1080/00221309.2012.695410

Swallow, K. M., Zacks, J. M., & Abrams, R. a. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology. General*, *138*(2), 236–57. doi:10.1037/a0015631

Underwood, G., & Swain, R. A. (1973). Selectivity of attention and the perception of duration. *Perception*, *2*(1), 101–105. doi:10.1068/p020101

van Wassenhove, V., Wittmann, M., Craig, A. D. B., & Paulus, M. P. (2011). Psychological and neural mechanisms of subjective time dilation. *Frontiers in Neuroscience*. doi:10.3389/fnins.2011.00056

Wittmann, M. (2013). The inner sense of time: how the brain creates a representation of duration. *Nature Reviews. Neuroscience*, *14*, 217–23. doi:10.1038/nrn3452

Wittmann, M., van Wassenhove, V., Craig, a D. B., & Paulus, M. P. (2010). The neural substrates of subjective time dilation. *Frontiers in Human Neuroscience*, *4*(February), 2. doi:10.3389/neuro.09.002.2010

Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: An emotional binding account. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2015.02.009

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–93. doi:10.1037/0033-2909.133.2.273

Zakay, D. (1992). The role of attention in children's time perception. *Journal of Experimental Child Psychology*, *54*(3), 355–371. doi:10.1016/0022-0965(92)90025-2

Zakay, D. (1998). Attention allocation policy influences prospective timing. *Psychonomic Bulletin & Review*, *5*(1), 114–118. doi:10.3758/bf03209465

Zakay, D., & Block, R. A. (1997). Temporal cognition. *Current Directions in Psychological Science*, *6*(1), 12–16. doi:10.1111/1467-8721.ep11512604

Zakay, D., & Block, R. A. (2004). Prospective and retrospective duration judgments: An executive-control perspective. In *Acta Neurobiologiae Experimentalis* (Vol. 64, pp. 319–328).

Zakay, D., Tsal, Y., Moses, M., & Shahar, I. (1994). The role of segmentation in prospective and retrospective time estimation processes. *Memory & Cognition*, *22*(3), 344–351. doi:10.3758/BF03200861

Zalla, T., Verlut, I., Franck, N., Puzenat, D., & Sirigu, A. (2004). Perception of dynamic action in patients with schizophrenia. *Psychiatry Research*, *128*(1), 39–51. doi:10.1016/j.psychres.2003.12.026

Zwaan, R. A. (1996). Processing Narrative Time Shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1196–1207. doi:http://dx.doi.org/10.1037/0278-7393.22.5.1196

Study 1: Neural Pattern Change During Encoding of a Narrative Predicts Retrospective Duration Estimates

Abstract

What mechanisms support our ability to estimate durations on the order of minutes? Behavioral studies in humans have shown that changes in contextual features lead to overestimation of past durations. Based on evidence that the medial temporal lobes and prefrontal cortex represent contextual features, we related the degree of fMRI pattern change in these regions with people's subsequent duration estimates. After listening to a radio story in the scanner, participants were asked how much time had elapsed between pairs of clips from the story. Our ROI analyses found that duration estimates were correlated with the neural pattern distance between two clips at encoding in the right entorhinal cortex. Moreover, whole-brain searchlight analyses revealed a cluster spanning the right anterior temporal lobe. Our findings provide convergent support for the hypothesis that retrospective time judgments are driven by "drift" in contextual representations supported by these regions.

Introduction

Imagine that you are at the bus stop when you run into a colleague and the two of you become engrossed in a conversation about memory research. After a few minutes, you realize that the bus still has not arrived. Without looking at your watch, you have some sense of how long you have been waiting. Where does this intuition come from?

Estimation of durations lasting a few seconds has been probed in the neuroimaging, neuropsychology and neuropharmacology literatures (see Wittmann, 2013, for a review). On the other hand, the neural mechanisms underlying time perception on the scale of minutes have remained unexplored. This is particularly true of *retrospective* judgments, where individuals experience an interval without paying attention to time and must subsequently estimate the interval's duration. In such cases, individuals must rely on information stored in memory to estimate duration. How is this accomplished?

Memory scholars have long posited that the same contextual cues that help us to retrieve an item from memory can also help us determine its recency. According to extant theories of context and memory (see Manning, Kahana, & Norman, 2014, for a review), *mental context* refers to aspects of our mental state that tend to persist over a relatively long time scale; this encompasses our representation of slowly-changing aspects of the external world (e.g., what room we are in) as well as other slowly-changing aspects of our internal mental state (e.g., our current plans). Crucially, these theories posit that slowly-changing contextual features can be episodically associated with more quickly-changing aspects of the world (e.g., stimuli that appear at a particular moment in time; Mensink & Raaijmakers, 1988; Howard & Kahana, 2002).

Bower (1972) first proposed that we could determine how long ago an item occurred by comparing our current context with the context associated with the remembered item. The similarity of these two context representations would reflect their temporal distance, with more similar representations associated with events that happened closer together in time. Thus, a slowly varying mental context could serve as a temporal tag (Polyn & Kahana, 2008). In parallel, researchers in the domain of retrospective time estimation have shown that the degree of context change is a better predictor of duration judgments than alternative explanations, such as the number of items remembered from the interval (Block & Reed, 1978; Block, 1990, 1992). Indeed, changes in task processing (Block & Reed, 1978; Sahakyan & Smith, 2014), environmental context (Block, 1982), and emotions (Pollatos, Laubrock, & Wittmann, 2014), as well as event boundaries (Poynter, 1983; Zakay, Tsal, Moses, & Shahar, 1994; Faber & Gennari, 2015), lead to overestimation of past durations.

In our study, we set out to obtain neural evidence in support of the hypothesis that mental context change drives duration estimates. Specifically, we hypothesized that, in brain regions representing mental context, the degree of neural pattern change between two events (operationalized as change in multi-voxel patterns of fMRI activity) should predict participants' estimates of how much time passed between those events.

Extensive prior work has implicated the medial temporal lobe (MTL) and lateral prefrontal cortex (PFC) in representing contextual information (Polyn & Kahana, 2008; for reviews of MTL contributions to representing context, see Eichenbaum, Yonelinas, & Ranganath, 2007, and Ritchey & Ranganath, 2012; for related computational modeling work, see Howard & Eichenbaum, 2013). In keeping with our hypothesis, multiple studies have obtained evidence

linking neural pattern change in these regions to temporal memory judgments. Manns, Howard, & Eichenbaum (2007) recorded from rat hippocampus during an odor memory task; they found that greater change in hippocampal activity patterns between two stimuli predicted better memory for the order in which the stimuli occurred. In the human neuroimaging literature, Jenkins & Ranganath (2010) found that the degree to which activity patterns in rostrolateral prefrontal cortex changed during the encoding of a stimulus predicted better memory for the temporal position of that stimulus in the experiment. Jenkins & Ranganath (2016) also showed that greater pattern distance between two stimuli at encoding in the hippocampus, medial and anterior prefrontal cortex predicted better order memory. Only one study has directly related neural pattern drift to judgments of elapsed time in humans: Ezzyat & Davachi (2014) found that patterns of fMRI activity in left hippocampus were more similar for pairs of stimuli that were later estimated to have occurred closer together in time, despite equivalent time passage between all pairs (a little less than a minute).

While the Ezzyat & Davachi (2014) study provides support for our hypothesis, it has some limitations. First, in Ezzyat & Davachi (2014), participants estimated the temporal distance of stimuli that were linked to their contexts in an artificial way (by placing pictures of objects or famous faces on unrelated scene backgrounds); it is unclear whether these results will generalize to more naturalistic situations where events are linked through a narrative. Second, since participants performed the temporal memory test after each encoding run, they were not entirely naïve to the manipulation. Knowing that they would have to estimate durations between stimuli could have changed participants' strategy and enhanced their attention to time (for evidence that estimating time prospectively engages different mechanisms, see Hicks,

Miller, & Kinsbourne, 1976, and Zakay & Block, 2004). In the current study, we sought to address the above issues by eliciting temporal distance judgments for pairs of events that had occurred several minutes apart and that were embedded in the context of a rich naturalistic story; participants listened to the entire story before being informed about the temporal judgment task.

Based on the studies reviewed above, we predicted that neural pattern drift in medial temporal and lateral prefrontal regions might support duration estimation. In our study, we examined these regions of interest (ROIs), as well as a broader set of regions that have been implicated in fMRI studies of time estimation, including the inferior parietal cortex, putamen, insula and frontal operculum (see ***Box 1*** for a review). In addition to the ROI analysis, which examined activity patterns in masks that were anatomically defined, we performed a searchlight analysis, which examined activity patterns within small cubes over the whole brain.

Participants were scanned while they listened to a 25-minute science fiction radio story. Outside the scanner, they were surprised with a time perception test, in which they had to estimate how much time had passed between pairs of auditory clips from the story. Controlling for objective time, we found that the degree of neural pattern distance between two clips at the time of encoding predicted how much time an individual would later estimate passed between them. The effect was significant in the right entorhinal cortex ROI. Extending the anatomical analysis to all masks in cortex revealed an additional effect in the left caudal anterior cingulate cortex (ACC). Moreover, whole-brain searchlight analyses yielded significant clusters spanning the right anterior temporal lobe. Our results suggest that patterns of neural

activity in these regions may carry contextual information that helps us make retrospective time judgments on the order of minutes.

Box 1. fMRI literature on prospective time estimation

As noted in the main text, only one study (Ezzyat & Davachi, 2014) has used fMRI to study retrospective estimation of time intervals lasting more than a few seconds. The vast majority of fMRI studies of time estimation have used prospective tasks, in which participants are asked to deliberately track the duration of a short stimulus or compare the duration of two stimuli. Such studies have repeatedly shown that activity in the putamen, insula, inferior frontal cortex (frontal operculum), and inferior parietal cortex increases as participants pay more attention to the duration of stimuli, as opposed to another time-varying attribute (Coull, 2004; Coull, Vidal, Nazarian, & Macar, 2004; Livesey, Wall, & Smith, 2007; Wiener, Turkeltaub, & Coslett, 2010; Wittmann, Simmons, Aron, & Paulus, 2010). Moreover, Dirnberger et al. (2012) showed that greater activity in the putamen and insula during encoding of aversive emotional pictures predicted better subsequent memory for those pictures, but only when their duration was overestimated relative to neutral images. This suggests that the putamen and insula might mediate the relationship between enhanced processing for emotional stimuli and subjective time dilation. Given the established role of these regions in time processing (albeit of a different sort) we included these regions in the set of a priori ROIs for our main fMRI analysis.

Results

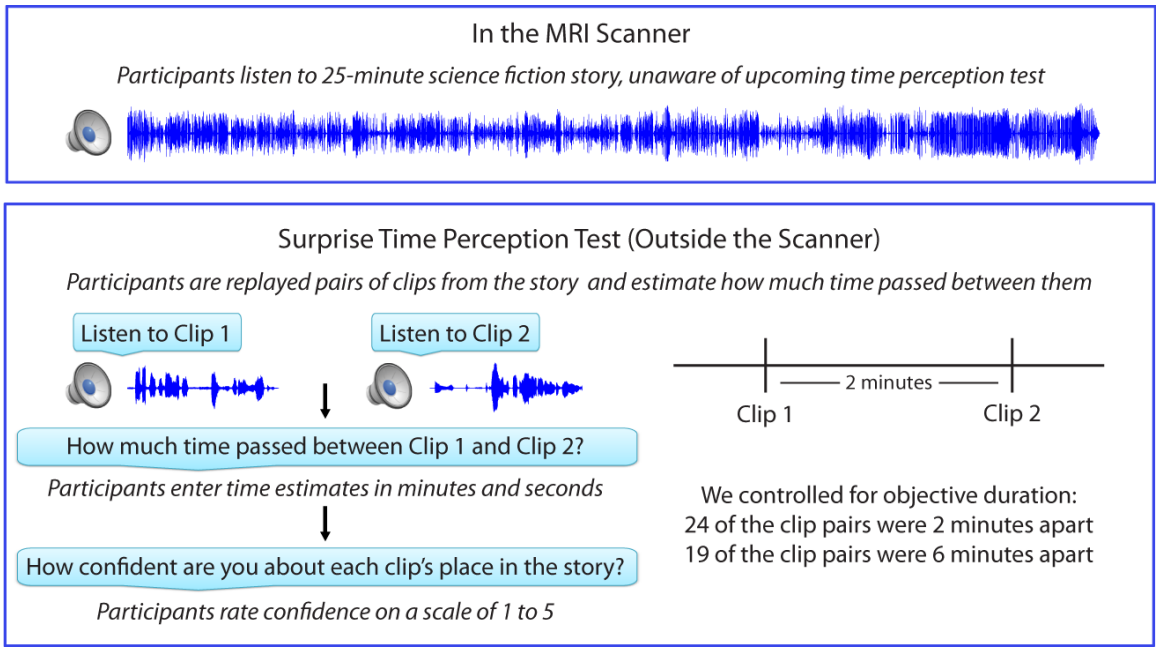


Figure 1 Experimental design

Behavioral Results

Participants were sensitive to the duration of story intervals

Figure 1 shows the experimental design, which consisted of an fMRI session, followed immediately by a behavioral session. After listening to a 25-minute radio story in the scanner, participants were asked how much time had passed between 43 pairs of clips from the story. In actuality, 24 of the clip pairs had been presented 2 minutes apart in the story, while 19 of the clip pairs had been presented 6 minutes apart in the story (participants were not informed of this). Participants were able to estimate the duration of experienced minutes-long intervals far

above chance, albeit with substantial intra- and inter-individual variability. On average, across participants, the 6-minute intervals (*M*=5.70 min, *SD*=3.06) were judged to be significantly longer than the 2-minute intervals (*M*=3.69 min, *SD*=1.96), $t(17) = 5.20$, $p = 0.00007$ (see ***Figure 2 A***).

As described in the Methods (see *Removing low-confidence intervals*), participants also provided confidence ratings reflecting their certainty about each clip's place in the story. Based on this measure, we grouped each participant's duration estimates into high-confidence and low-confidence intervals. To verify that participants were better at distinguishing 6-minute intervals from 2-minute intervals when they were confident, we calculated the difference between the mean duration estimates for 6-minute intervals and the mean duration estimates for 2-minute intervals for every participant. The difference score was significantly higher for high-confidence intervals (*M*=2.43, *SD*=1.82) than for all intervals (*M*=2.01, *SD*=1.64), $t(17)=2.33$, $p=0.0324$. Thus, participants were significantly more accurate at estimating an interval's duration when they confidently remembered the temporal position of both clips delimiting that interval in the story (see ***Figure 2 B***).

For a given interval duration, some intervals were consistently judged to be longer than other intervals across participants, although the actual amount of elapsed time was held constant. To test the reliability of duration estimates across participants, we split the subjects randomly into two groups, averaged the duration estimates within each group, and correlated the two averages with each other. We repeated this procedure 1000 times to ensure that we sampled a variety of group splits. The average correlation between the two groups was 0.64 (SD=0.09) for 2-minute intervals and 0.54 (SD=0.15) for 6-minute intervals (see **Figure 2 -**

Supplement 1). This analysis suggests that features of the story made some intervals appear consistently shorter and other intervals appear consistently longer across participants.

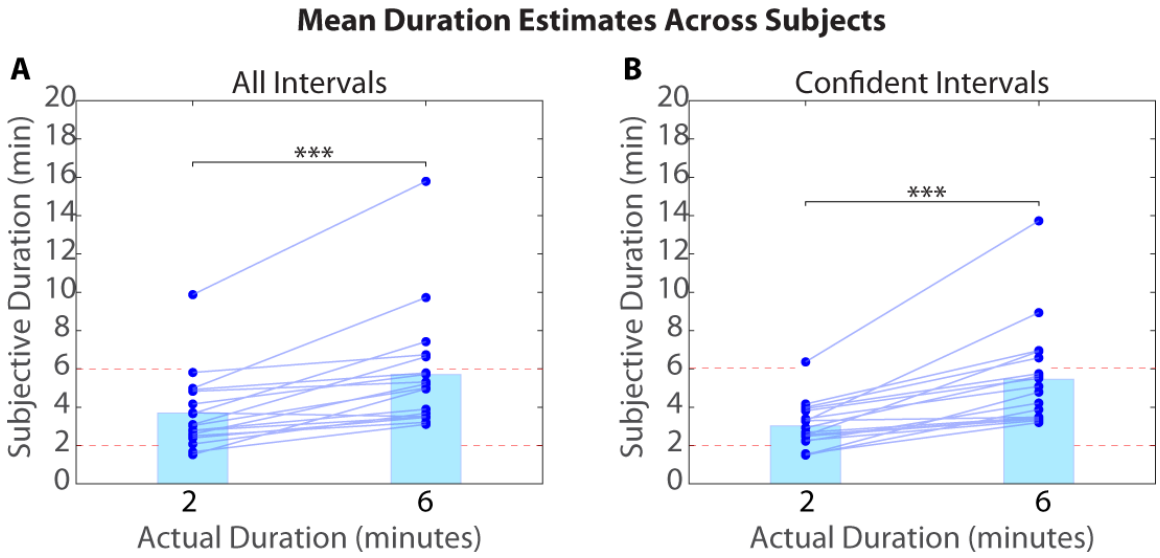
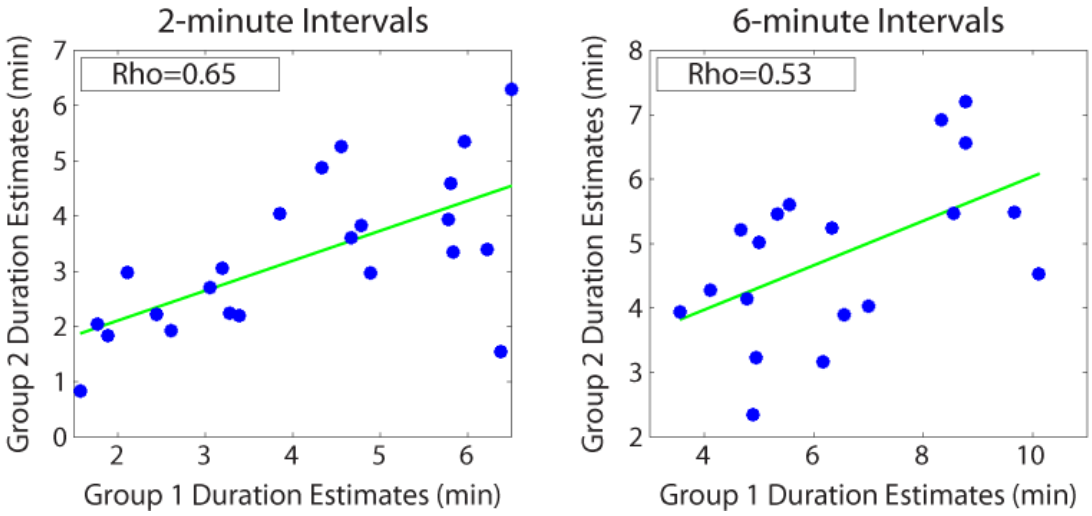


Figure 2 Mean duration estimates for all intervals (A) and confident intervals (B) as a function of their actual duration. Each blue circle represents the mean duration estimate for an individual participant within a given interval duration (2 or 6 minutes). The blue bar heights represent the global means for 2 and 6-minute intervals across intervals and participants.

Figure 2 – Supplement 1: Reliability of duration estimates across participants



Reliability of duration estimates across participants. Between-group correlations were obtained by splitting the participants randomly into two equal groups and averaging the duration estimates for each interval (across participants) within a group. Each dot in the scatterplot represents a particular temporal interval; its *x* and *y* coordinates indicate the mean estimated duration of that interval for Group 1 and Group 2 participants, respectively. We repeated this procedure 1000 times to ensure that we sampled a variety of group splits. The average correlation between the two groups was 0.64 (*SD*=0.09) for 2-minute intervals and 0.54 (*SD*=0.15) for 6-minute intervals. The above plot shows the grouping that was most representative of the mean.

Duration Estimates Are Influenced by Memory of the Story

We found that participants estimated six-minute intervals to be significantly longer than two-minute intervals (**Figure 2**), and that some intervals in the story tended to be systematically over-estimated by participants (**Figure 2 – Supplement 1**). However, it is possible that participants could judge the temporal distance between two clips purely based on the similarity between them (e.g. Are the same characters speaking? Is the background music the same? Is the topic of conversation similar?)

To ensure that participants were using their memory of the story to judge temporal distance, we ran a control experiment in which 17 participants who had never heard the story were given the exact same memory test. They were asked to try to estimate the amount of time that had elapsed between each pair of clips during the original telling of the story. During debriefing, participants reported making duration estimates based on the perceptual and semantic similarity between the two clips (e.g., which character voices were present, which background music was playing, the topic of conversation).

We found that naïve participants estimated 6-minute intervals (M=6.21 min, SD=1.91) to be longer than 2-minute intervals (M=5.63 min, SD=1.74; $t(16)=2.62$, p=0.019), suggesting that the similarity between two clips carried some information about the temporal distance between them. However, naïve participants were significantly less accurate at distinguishing 6-minute intervals from 2-minute intervals than our original participants who had heard the story. To quantify this, we calculated the difference between the mean duration estimates for 6-minute intervals and the mean duration estimates for 2-minute intervals for every participant (exactly

as above). The difference score was significantly higher for our original participants (M=2.01 min, SD=1.64 min) than for naïve participants (M=0.59 min, SD=0.91 min), $t(26.86)= -3.22$, $p<0.005$. Thus, having memory of the story enabled our participants to estimate durations with significantly higher accuracy.

We hypothesized that both our original participants and the naïve participants would use consistent strategies to estimate the temporal distance between two clips, but that these strategies would differ across groups. If this is the case, duration estimates should be more correlated across participants within groups than across participants between groups. The inter-subject correlation (see *Methods*) in duration estimates was as strong for naïve participants (M=0.43, SD=0.18, 95% CI [0.40, 0.56]) as for our original participants (M=0.43, SD=0.25, 95% CI=[0.37, 0.58]), suggesting that both groups used a consistent strategy to estimate the distance between two clips. When we correlated duration estimates from our original group of participants with those of our naïve participants, we found that the between-group correlations (M=0.18, SD=0.22, 95% CI=[0.04, 0.28]) were significantly above 0, suggesting that a component of the original duration estimates was influenced by the similarity in content between clips. However, the between-group correlations were significantly lower than the within-group correlations ($p<0.0001$, as assessed by a permutation test described in the *Methods*). In other words, there is a reliable component of our original participants' behavior that cannot be captured by accounting for the perceptual and semantic similarity between clips. In summary, having memory of the story induced a qualitatively different pattern of behavior and produced significantly more accurate duration estimates.

Correlation between number of event boundaries and duration estimates

To gain additional evidence that duration estimates were related to contextual change, we looked at the correlation between estimated duration and the number of event boundaries in the interval between the clips. The number of intervening event boundaries can be viewed as a proxy for contextual change, insofar as event boundaries often encompass changes in scene, characters and conversation topic (Kurby & Zacks, 2008; Zacks, Speer, & Reynolds, 2009). As reviewed in the *Introduction*, numerous studies have found a relationship between changes in contextual features during an interval and duration estimates for that interval.

A separate group of participants (n=9) listened to the story and was asked to press a button every time they felt an event boundary was occurring. These data were then averaged across participants to obtain the mean number of event boundaries inside each two-minute interval. We found that the mean number of boundaries in an interval was significantly correlated with the mean duration estimates from our original experiment (r=0.49, 95% CI [0.27, 0.57]; ***Figure 3***). This suggests that our participants' retrospective duration estimates were influenced by the number of contextual changes that had occurred during an interval.

However, it is important to note that the number of event boundaries between two clips also influences the perceptual and semantic similarity between them (e.g., clips from the same scene might sound more similar than clips from different scenes). Thus, our participants' duration estimates could correlate with the number of event boundaries, even if they were basing their estimates purely on the perceptual similarity between clips. To explore this

possibility, we tested whether the number of event boundaries would correlate with duration estimates from naïve participants, who could *only* judge temporal distance based on the similarity between clips, given that they had never heard the story.

Importantly, we found that the number of event boundaries in an interval did not significantly correlate with duration estimates of naïve participants (r=0.09, 95% CI [-0.05, 0.21]; ***Figure 3***). Of course, we cannot definitely prove the null hypothesis that naïve estimates do not at all correlate with the number of event boundaries. However, the correlation between the number of boundaries and duration estimates was significantly higher for our original participants than for naïve participants (r_{diff} = 0.40, 95% CI [0.15 0.56]). In other words, duration estimates from participants who remembered the story were significantly more correlated with the number of contextual changes between two clips than duration estimates from participants who were judging temporal distance based merely on the similarity between the two clips. This suggests that the number of event boundaries carries information about temporal context that is not contained within the clips alone, and that our original participants' estimates were influenced by their memory of this contextual information.

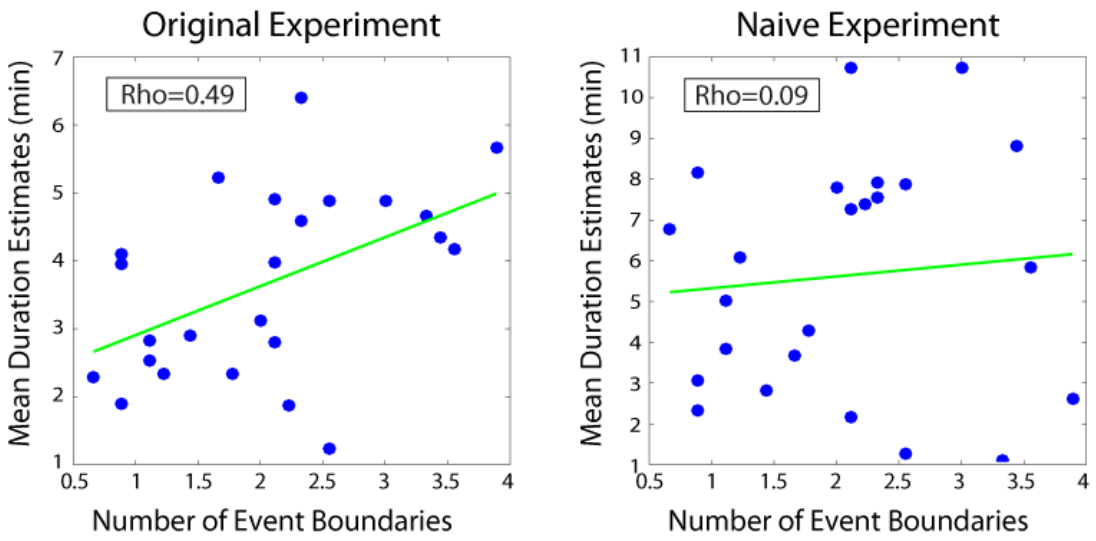


Figure 3 Mean duration estimates for 2-minute intervals as a function of the number of event boundaries in each interval. The number of event boundaries in an interval predicted retrospective duration estimates in our original experiment (left panel), but did not predict duration estimates of naïve participants (right panel) who had never heard the story. This suggests that the number of contextual changes between two clips influenced temporal distance judgments only when the content of the story between the two clips could be recalled.

fMRI Results

We tested whether BOLD pattern change between two clips correlated with temporal distance estimates, using both ROI and whole-brain searchlight analyses. Each type of analysis was performed both within-participants across intervals and within-intervals across participants.

In the within-participant analysis, we correlated each participant's duration estimates with that participant's neural pattern distances (see ***Within-Participant Correlation between Pattern Change and Duration Estimates*** and ***Within-Participant Whole-brain Searchlight***). In the within-interval analysis, we correlated individual differences in subjective duration for a given interval with individual differences in neural pattern distance for that interval (see ***Within-Interval Correlation between Pattern Change and Duration Estimates*** and ***Within-Interval Whole-brain Searchlight***). The two versions of each analysis were performed in order to rule out the possibility that our effects were driven either by participant or interval random effects. In particular, we were concerned that correlations between neural pattern distance and behavior could reflect sensitivity to perceptual or semantic features of the clips (i.e., clip pairs with similar perceptual/semantic features might be associated with shorter duration estimates and greater neural similarity, relative to clip pairs with more dissimilar features). The within-interval analysis addresses this concern by holding clip identity constant.

Next, we fit a mixed-effects model for each ROI (see ***Mixed-Effects Model Accounting for Naïve Duration Estimates***), in which we estimated whether pattern distance in that ROI could predict duration estimates, even when accounting for participant random effects, item

(interval) random effects, as well as naïve duration estimates (which are a proxy for the perceptual and semantic similarity between two clips, see ***Behavioral Results***).

Finally, we discuss the brain regions that showed significant effects across all analyses (see ***Comparing Results from ROI and Searchlight Analyses***).

As noted in the *Methods*, the ROI and searchlight analyses were conducted only on high-confidence 2-minute intervals. 6-minute intervals were excluded from the fMRI analysis, since we could not successfully dissociate neural pattern change at this timescale from low-frequency scanner noise (see ***Methodological challenges with analyzing pattern distance over long time scales*** in the *Methods*).

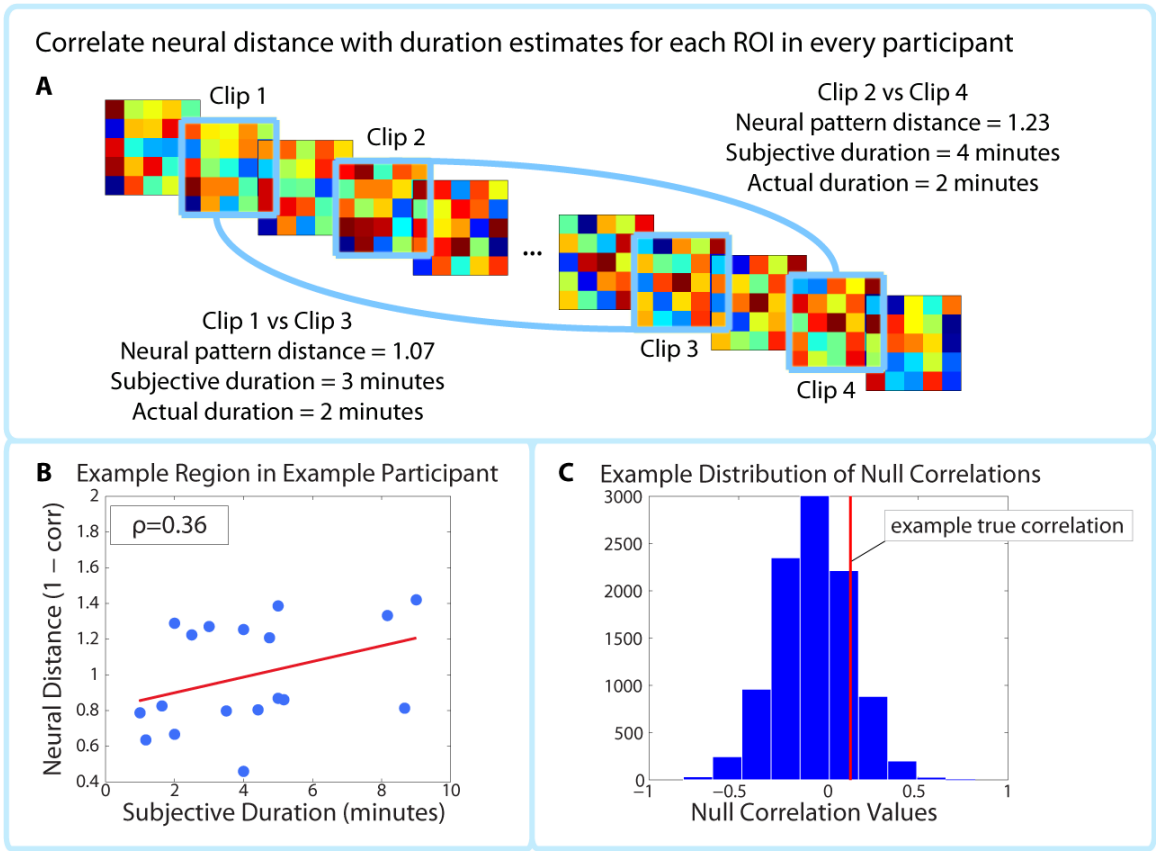
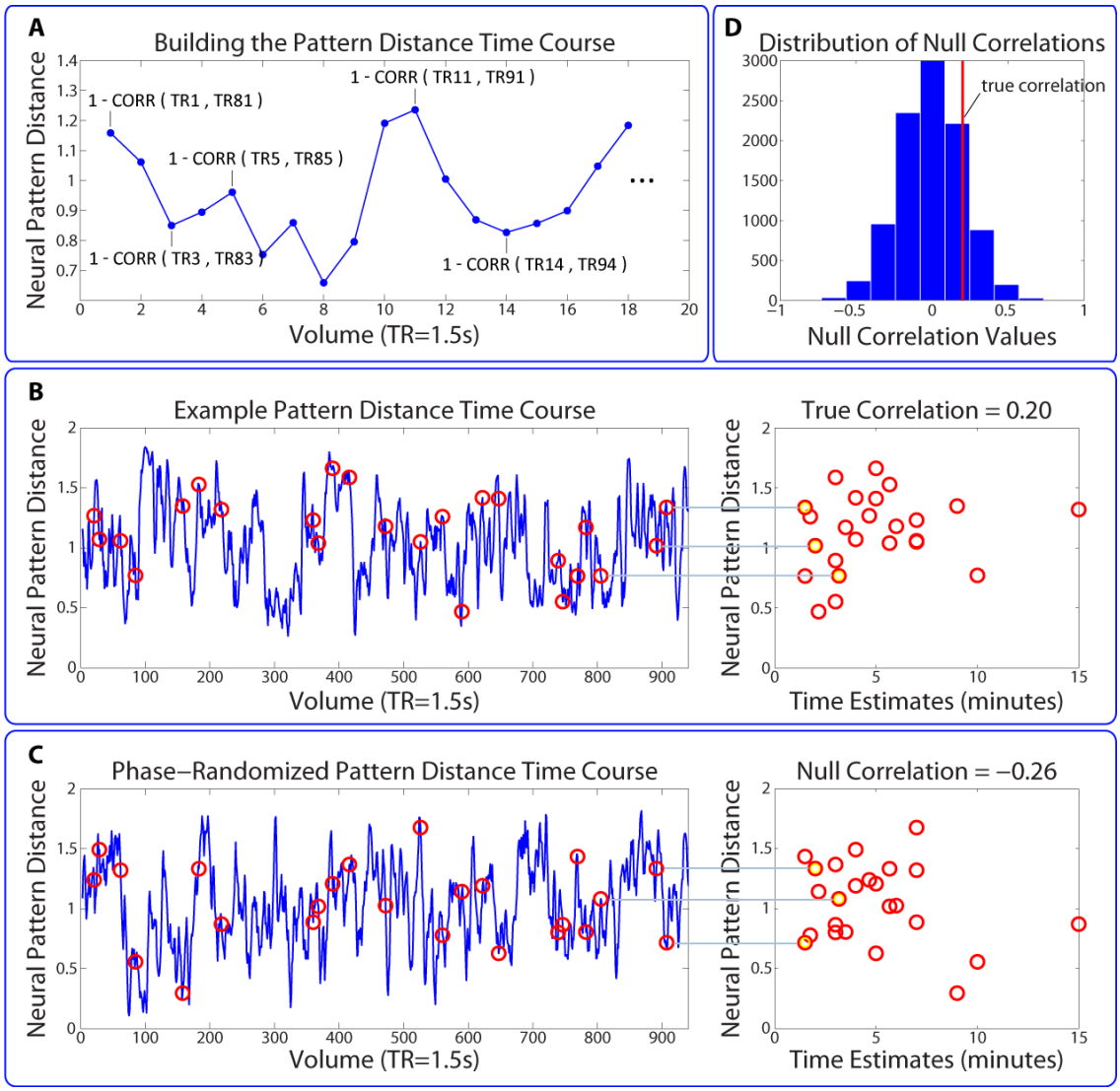


Figure 4 Correlating pattern distance with duration estimates within participants. For each ROI in each participant, the pattern distance between each pair of clips at encoding was correlated with the participant's retrospective duration estimate (A-B). The top panel (A) shows two example intervals. The neural distance (1-Pearson's r) between clips 2 and 4 (second interval) is greater than the neural distance between clips 1 and 3 (first interval), as is the subjective duration estimate. (B) shows the correlation between neural distance and duration estimates in a hypothetical region and participant. (C) We used a permutation test to generate 10,000 surrogate pattern distance vectors (see Figure 3 - Supplement 1), which we then used to obtain a distribution of null correlations between neural distances and duration estimates. For each ROI in each participant, we calculated the z-scored correlation value, which reflects the strength of the empirical correlation relative to the distribution of null correlations. For each ROI, we performed a random effects t-test to assess whether the z-score was reliably positive across participants. P-values from this t-test were then subjected to multiple comparisons correction using False Discovery Rate (FDR).

Figure 4 – Supplement 1: Permutation test assessing the temporal specificity of correlations between pattern change and behavior



Permutation test assessing the temporal specificity of correlations between pattern change and behavior. This procedure is described in the *Methods* (see "Statistical analysis of correlations between pattern change and behavior"). (A,B) The time course of pattern change is constructed using the distance (1 - Pearson's *r*) between each pattern and the pattern 80 TRs (2 minutes) after it. As in the main analysis, we averaged over the 5 consecutive TRs surrounding each pattern (for simplicity, this is not shown in the above figure). (C) 10 000 surrogate pattern distance time courses are generated by randomizing the phases of the original time course, thus conserving the amplitude of each frequency component. (D) Surrogate pattern distances are correlated with time estimates, generating 10,000 null correlations. A Z-value for each ROI / searchlight in each participant is computed to compare the strength of the empirical correlation with the distribution of null correlations. The p-value for a given ROI is obtained using a right-tailed t-test on the Z-values across participants.

Anatomical ROI Analyses

We first tested whether pattern change in regions suggested by the literature to be important for representing temporal context (see *ROI Selection*) correlated with retrospective duration estimates. Anatomical ROIs were derived from FreeSurfer cortical parcellation (Desikan et al., 2006) and from a probabilistic MTL atlas (Hindy & Turk-Browne, 2015).

Within-Participant Correlation between Pattern Change and Duration Estimates

The within-participant analysis procedure is outlined in ***Figure 4***. We calculated the correlation between neural pattern distance and duration estimates within participants (***Figure 4 A***) in each of the 32 ROIs shown in ***Figure 5***. To assess the likelihood of obtaining a correlation of that magnitude by chance, we used a phase randomization procedure (described in *Methods*) to obtain 10 000 null correlations for each ROI in every participant. This enabled us to calculate a Z-value for every ROI in every participant, which reflects the strength of the actual correlation between pattern distance and duration estimates relative to the distribution of null correlations (***Figure 4 C***). Here we report the regions whose Z-values were consistently positive across participants, corrected for multiple comparisons using False Discovery Rate (FDR, Benjamini, Krieger, & Yekutieli, 2006).

Out of the regions selected a priori, the right entorhinal cortex and right pars orbitalis showed a significant positive correlation between pattern change and duration estimates for high-confidence 2-minute intervals (q<0.05). ***Figure 5*** shows the mean Z-values across

participants for all a priori ROIs (16 in each hemisphere), including lateral prefrontal regions (top panel A), medial temporal lobe regions, insula, putamen, and inferior parietal cortex (bottom panel B). While a large number of these regions had Z-values that were positive across participants (e.g., left hippocampus, left entorhinal cortex, right perirhinal cortex, right amygdala, bilateral insula, and right caudal middle frontal cortex, $p<0.05$ uncorrected), we report only those that survived FDR correction.

As part of an exploratory search, we also performed this analysis on the other brain regions derived from FreeSurfer cortical parcellation. This included the 16 ROIs mentioned above, in addition to regions in the occipital lobe, parietal lobe, medial prefrontal cortex, lateral temporal lobe, basal ganglia, thalamus and brainstem (the complete list of regions can be found in ***Figure 5 – source data 1***). Out of the 84 regions tested (42 in each hemisphere), the right entorhinal cortex, right pars orbitalis, and left caudal anterior cingulate cortex (ACC) showed significant positive correlations between pattern change and duration estimates ($q<0.1$). This suggests that the right entorhinal cortex and right pars orbitalis, which were part of our list of a priori ROIs, contained effects that were apparent even after whole-brain correction, and reveals an additional effect in the left caudal ACC that we had not anticipated. ***Figure 5 – Supplement 1*** displays the locations of these three regions in MNI space.

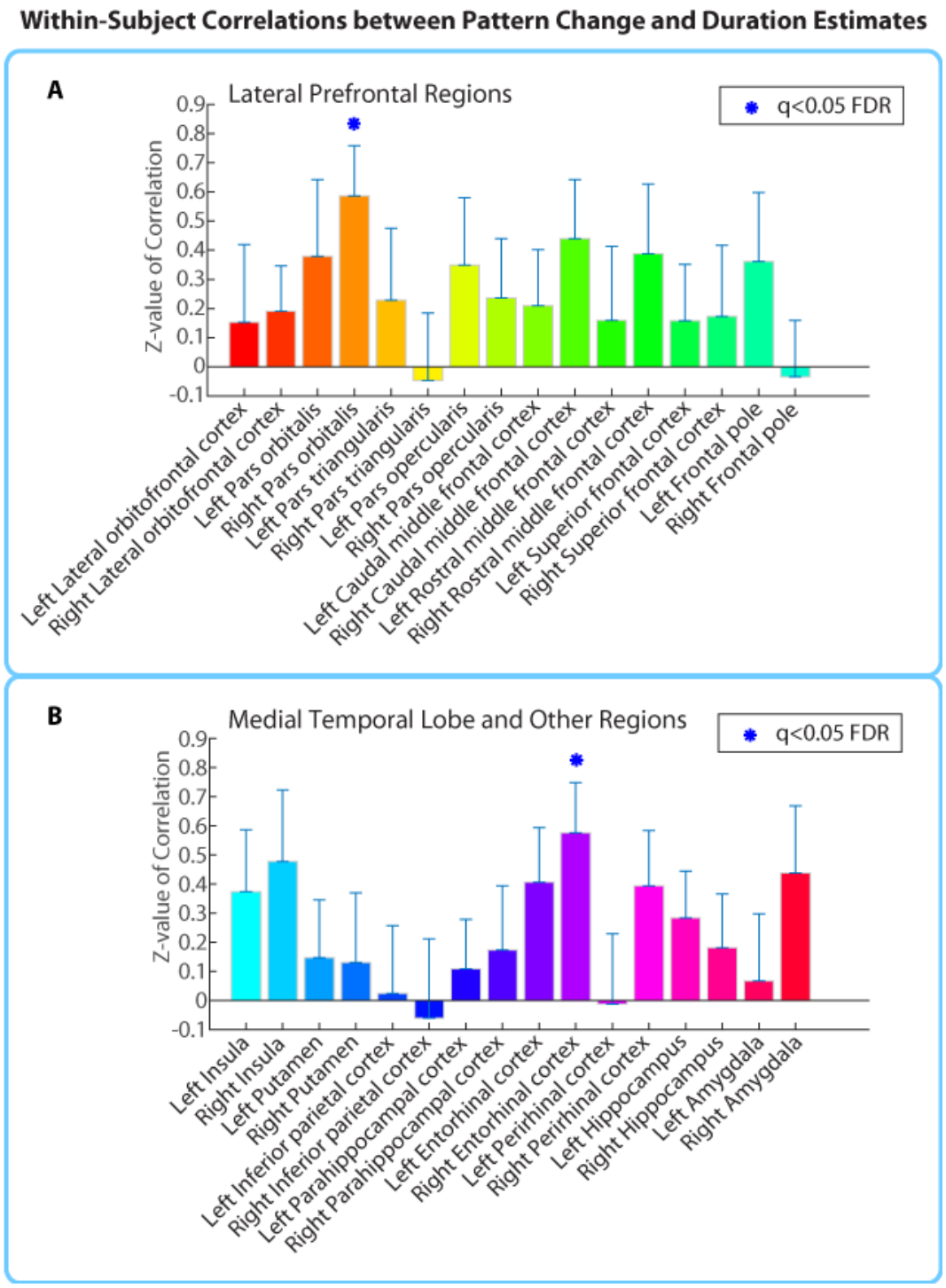
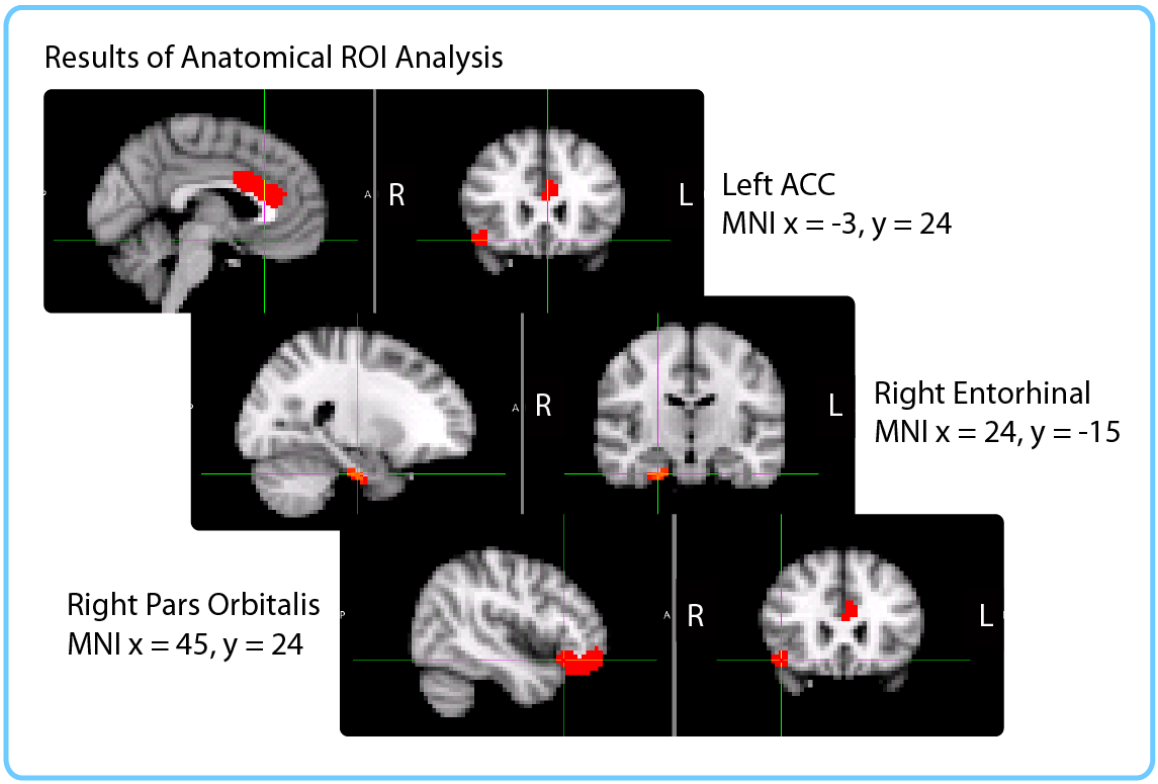


Figure 5 Within-participant ROI analysis: Mean Z-values (across all 18 participants) of correlations between pattern distance and duration estimates for the 16 a priori ROIs. Z-values were obtained from the phase randomization procedure and reflect the strength of the empirical correlation relative to the distribution of null correlations. Error bars represent standard errors of the mean. The blue dots over the right entorhinal cortex and right pars orbitalis indicate that these ROIs survived FDR correction at q<0.05.

Figure 5 – Supplement 1



Anatomical ROIs that showed a significant correlation between pattern change and duration estimates within participants, after whole-brain FDR correction. In red are regions with q<0.1: the right entorhinal cortex, right pars orbitalis and left caudal ACC. This analysis was performed in native space on participant-specific ROIs. ROIs were transformed from native functional space to MNI space for display purposes.

Within-Interval Correlation between Pattern Change and Duration Estimates

Above, in the within-participants analysis, we found that the neural pattern distance between two clips at encoding was correlated with retrospective duration judgments in the right entorhinal cortex, right pars orbitalis and left caudal ACC. However, in the *Behavioral Results*, we found that the perceptual and semantic similarity between two clips could explain some of the variance in subjective duration across intervals, even though it could not explain all

the variance. Thus, it is possible that neural pattern change in the regions we found correlates with the component of duration estimates that is driven by perceptual and semantic content, rather than the component that is driven by abstract, slowly varying contextual features.

To rule out this concern, we performed a within-interval (across participants) version of the ROI analysis. For each ROI, we correlated 1) duration estimates for a given interval across participants with 2) the neural pattern distances for that interval across participants; results were then aggregated across all 2-minute intervals. Rather than capturing variance within an individual across intervals of the story, this analysis captures variance across individuals for a given interval of the story. By performing the correlation within a given interval, we hold constant the perceptual and semantic content of the two clips and only leverage individual differences in how long the interval appeared retrospectively.

As described in the *Methods*, a permutation test was used to assess the statistical significance of each correlation. Duration estimates were scrambled across participants 10,000 times to obtain a distribution of null correlations for every interval in every ROI. This enabled us to calculate a Z-value, which reflects the strength of the actual correlation between pattern distance and duration estimates relative to the distribution of null correlations. Finally, a right-tailed t-test was performed to assess whether the Z-values for a region were reliably above 0 across intervals. The p-values from this t-test were subjected to multiple comparisons correction using FDR.

Out of the regions selected *a priori*, the right entorhinal cortex, right amygdala, and right insula showed a significant positive correlation between pattern change and duration estimates

for high-confidence 2-minute intervals (q<0.05). ***Figure 6*** shows the mean Z-values across intervals for all *a priori* ROIs (16 in each hemisphere).

Extending this analysis to the whole brain (same anatomical masks as in ***Figure 5 – source data 1***) revealed only the right entorhinal cortex (q<0.05), suggesting that the effect in this region was strong enough to survive whole-brain correction.

Importantly, the right entorhinal cortex is the only region with significant effects in both the within-interval analysis (Cohen's d = 0.83) and the within-participant analysis (Cohen's d = 0.79). If neural pattern distance between two clips in entorhinal cortex were driven solely by changes in clip content, we would have expected the correlation with duration estimates to be larger for the within-participant analysis (where story content differed across intervals) than for the within-interval analysis (where story content is held constant). The fact that the effect sizes are similar shows that perceptual or semantic differences in content between the two clips are not the main factor driving the correlation between duration estimates and neural pattern change in this region.

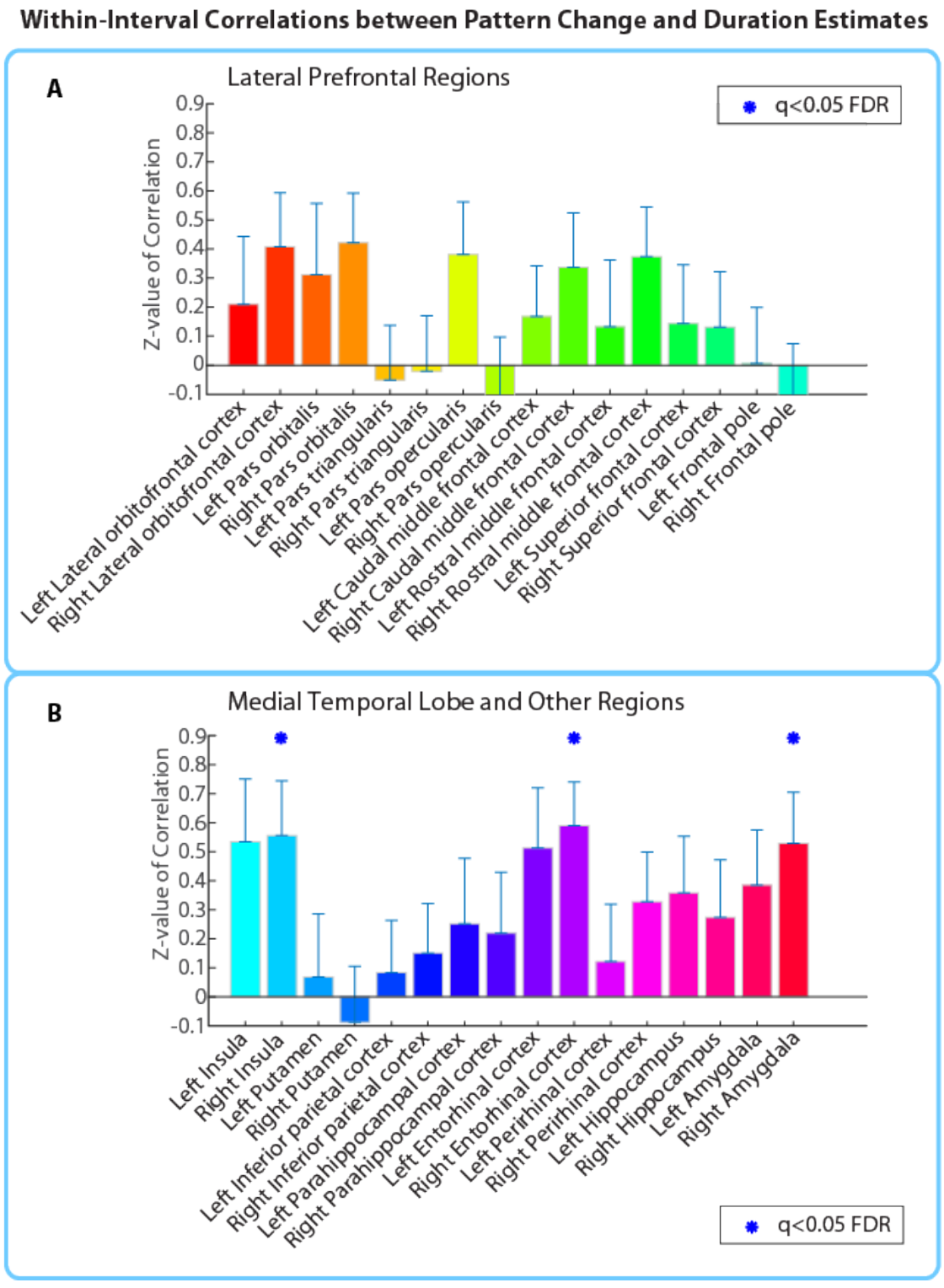


Figure 6 Within-interval ROI analysis: Mean Z-values (across all 2-minute intervals) of correlations between pattern distance and duration estimates for the 16 a priori ROIs. Error bars represent standard errors of the mean. Correlations between pattern change and duration estimates were performed across participants, separately for each interval.

Mixed-Effects Model Accounting for Naïve Duration Estimates

We analyzed our data using a hierarchical linear regression model (Gelman & Hill, 2006; see *Methods* for additional detail). This analysis estimates population-level effects of interest, while controlling for the possibility of individual variability between subjects and between clip pairs. In other words, this approach leverages the power of the within-interval analysis to control for the objective content similarity between two clips, while also taking into account variability in the effect across participants. In addition, we included the mean duration estimates from our naïve participants as a covariate in the model (see *Behavioral Results*). Since naïve participants had estimated the temporal distance between each pair of clips without hearing the story, this covariate is a further control for the inherent guessability of the temporal distance between two clips. Both controls strengthen our interpretation that the remaining effect of neural pattern distance on duration estimates is driven by the contextual dissimilarity (rather than perceptual or content dissimilarity) between two clips.

For each anatomical region derived from FreeSurfer and MTL segmentation (42 in each hemisphere), we fit a model where duration estimates were predicted by naïve duration estimates as well as the neural pattern distance in that region (see *Methods* for the complete formula). We then computed 95% confidence intervals of the fixed-effects parameter estimates using the asymptotic Gaussian approximation (see *Methods*).

The fixed effect of naïve estimates was positive in all models and its confidence intervals did not include zero in 80% of the models. This reproduced our finding that naïve duration estimates are correlated with the original duration estimates (see *Behavioral Results*),

suggesting that interval durations are partially guessable based on the similarity between clips. However, even under this control, the fixed effect of neural pattern distance in left caudal ACC and right entorhinal cortex exhibited confidence intervals that did not include zero (***Figure 7***). ***Figure 7 – source data 1*** contains the parameter estimates and 95% confidence intervals for all 84 anatomical regions.

Importantly, including the naïve duration estimates as a covariate in the model did not significantly weaken the relationship between neural pattern distance and duration estimates in these regions (though the effects were slightly lower numerically). ***Figure 7*** shows in green the 95% confidence intervals for the same ROIs when naïve duration estimates are excluded from the model.

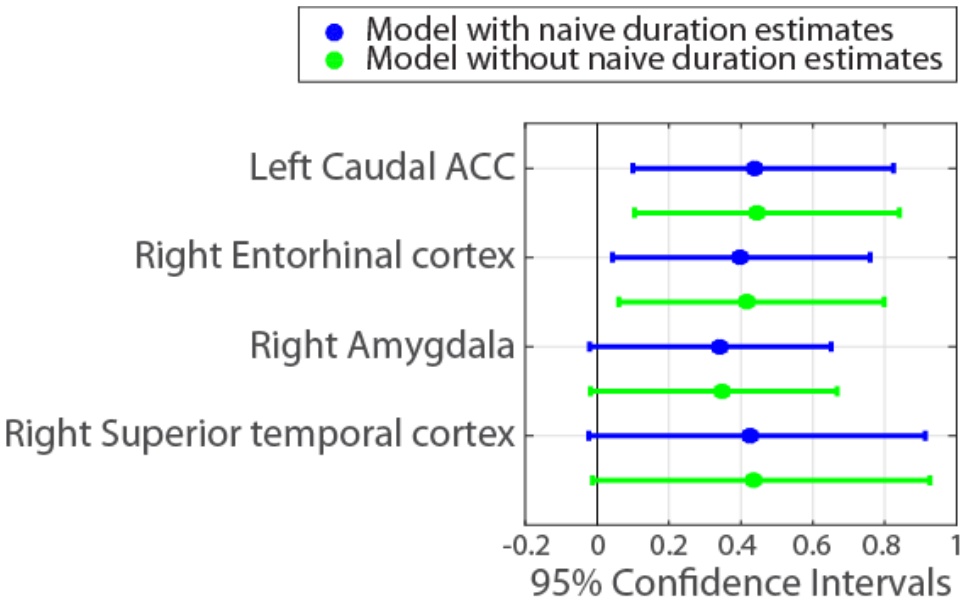


Figure 7 Parameter estimates and 95% confidence intervals for the fixed effect of neural pattern distance on duration estimates. We also included the right amygdala and right superior temporal cortex in the figure, because their confidence intervals did not include 0 when a slightly less conservative fitting procedure was used (see Methods).

Whole-Brain Searchlights

As with the Anatomical ROI analyses, both within-participant and within-interval analyses were performed for the Whole-Brain Searchlight analyses, in order to rule out the possibility that our effects were driven either by participant or interval random effects**.**

Within-participant Whole-brain Searchlight

We ran a cubic searchlight with 3x3x3 (27) voxels (972 mm^3) through the entire brain and tested for a correlation between pattern change and duration estimates in each searchlight. The same phase-randomization procedure that was used for the within-participant anatomical ROI analysis was also applied here; this procedure generates Z-values that reflect how likely we are to get this strong of a correlation by chance, given the frequency spectrum of the fMRI data. When excluding low-confidence intervals, we found a significant cluster in the right anterior temporal lobe (p=0.034, FWE-corrected; Center of Gravity MNI coordinates (x, y, z) in mm: [45.6, -5.53, -21.7]; cluster size=572 voxels in 3 mm MNI space). Small parts of the cluster also extended to the right posterior insula and right putamen (see ***Figure 8***).

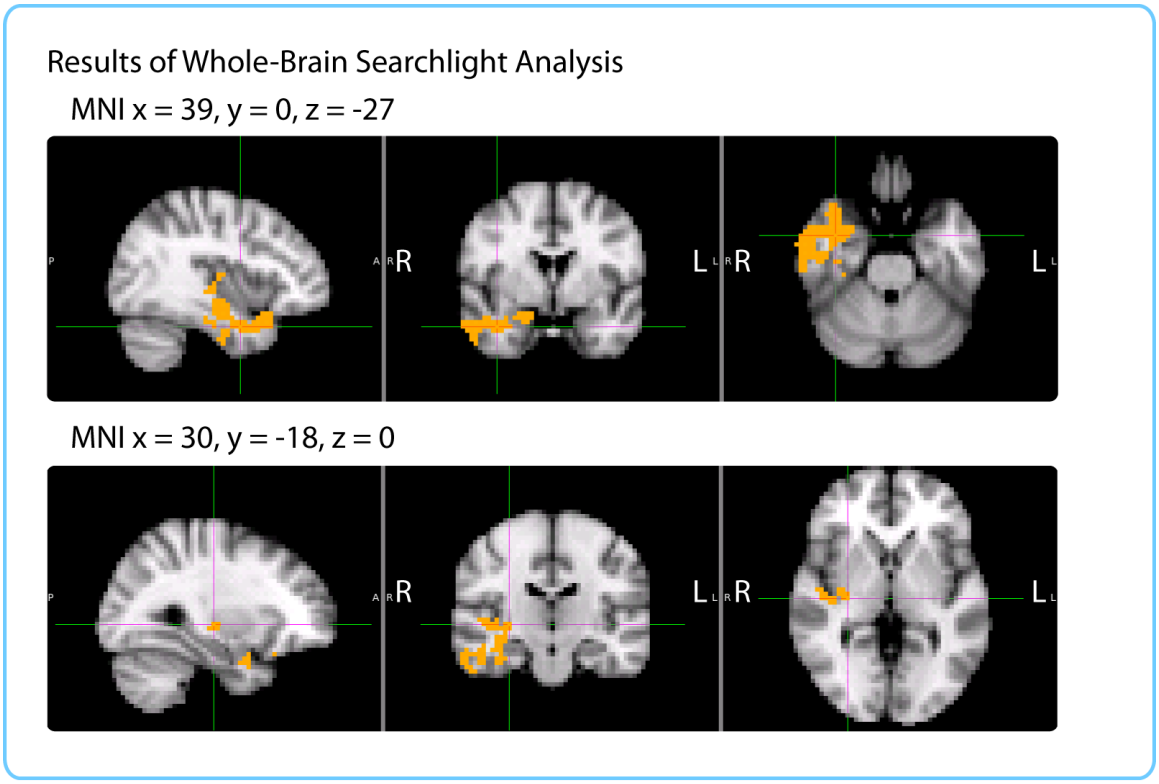


Figure 8 Results of within-participant whole-brain searchlight. Voxels in orange represent centers of searchlights that exhibited significant correlations between pattern change and duration estimates within participants across intervals ($p<0.05$, FWE). The significant cluster had peak MNI coordinates (in mm): x = 45.6, y = -5.53, z = -21.7.

Within-interval Whole-brain Searchlight

We also ran a searchlight version of the within-interval analysis. In order to match searchlights across participants, functional data were transformed to 3mm MNI space. Since this transformation approximately doubles the number of brain voxels, we ran cubic searchlights of radius 2 with 5x5x5 (125) voxels through the entire brain.

As with the ROI analysis, this analysis was performed on high-confidence duration estimates. For each interval, we only included participants who had confidently recollected the temporal position of the two clips delimiting that interval.

To assess the significance of each correlation score, we used the same permutation test as for the ROI analysis. Duration estimates were scrambled across participants 10,000 times to obtain a distribution of null correlations, and Z-values were calculated for each interval. We thus obtained a brain map of Z-values for each of the 24 intervals, and FSL's *randomise* function was used to control the Family-wise error rate, as above.

Similarly to the within-participant searchlight, we found a significant cluster in the right anterior temporal lobe (p=0.019, FWE-corrected; Center of Gravity MNI coordinates (x, y, z) in mm: [32.1, -10.2, -18.7]; cluster size=535 voxels in 3 mm MNI space). The cluster extended from the right parahippocampal gyrus, hippocampus and amygdala medially to the middle temporal gyrus and temporal pole laterally (see ***Figure 9***).

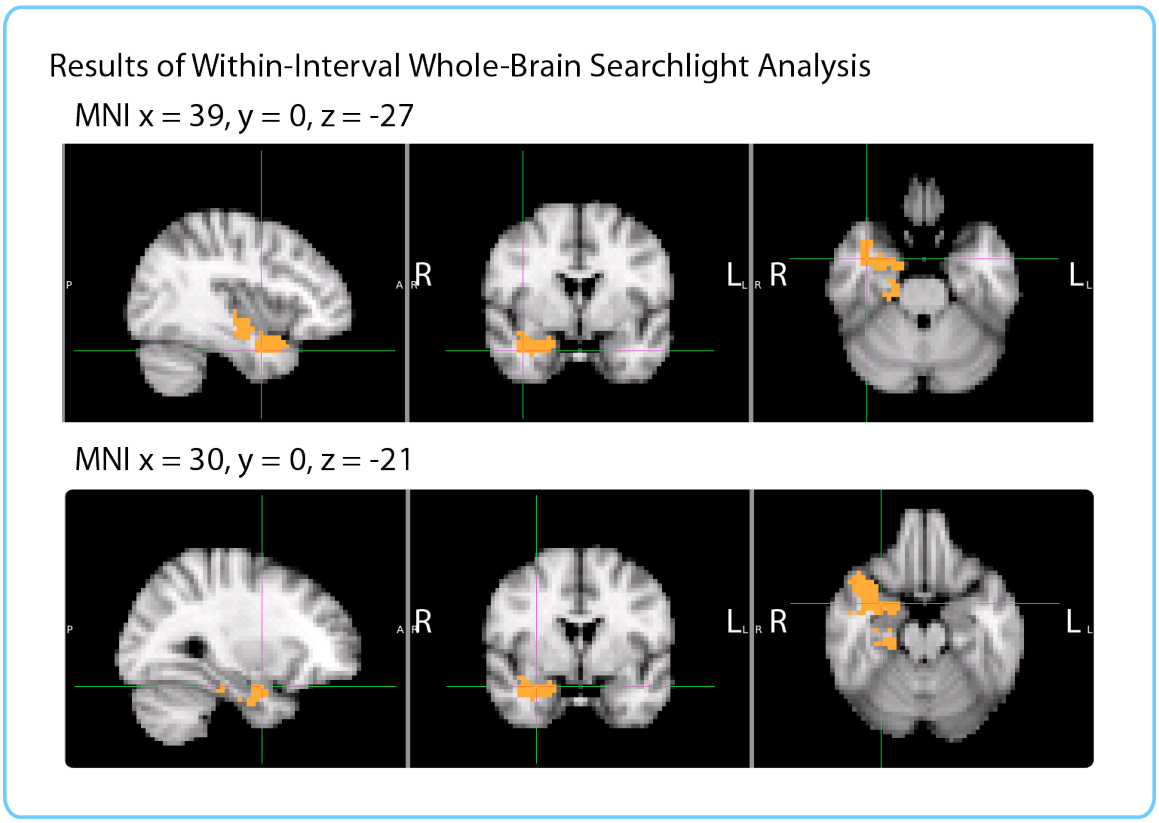


Figure 9 Results of within-interval whole-brain searchlight. Voxels in orange represent centers of searchlights that exhibited significant correlations between pattern change and duration estimates across participants ($p<0.05$, FWE). The significant cluster had center of gravity MNI coordinates (in mm): x =32.1, y = -10.2, z = -18.7.

Comparing Results from ROI and Searchlight Analyses

The within-participant ROI analysis revealed significant effects in the right entorhinal cortex, right pars orbitalis and left caudal ACC. The within-interval ROI analysis revealed significant effects in the right entorhinal cortex, right amygdala and right insula. The mixed-effects ROI analysis showed that the right entorhinal cortex and left caudal ACC had confidence intervals above 0, even when naïve duration estimates were accounted for. Both the within-participant and within-interval searchlights revealed significant clusters in the right anterior temporal lobe.

Figure 10 enables a comparison of the two searchlight analyses; the right entorhinal cortex ROI that emerged in all three ROI analyses is also overlaid. The within-interval searchlight cluster was located more medially than the within-participant searchlight cluster, though the two overlapped in the right amygdala, right temporal pole, and the cerebral white matter of the right anterior temporal lobe. Moreover, the within-interval searchlight cluster overlapped with the right entorhinal cortex ROI (see green voxels, ***Figure 10***).

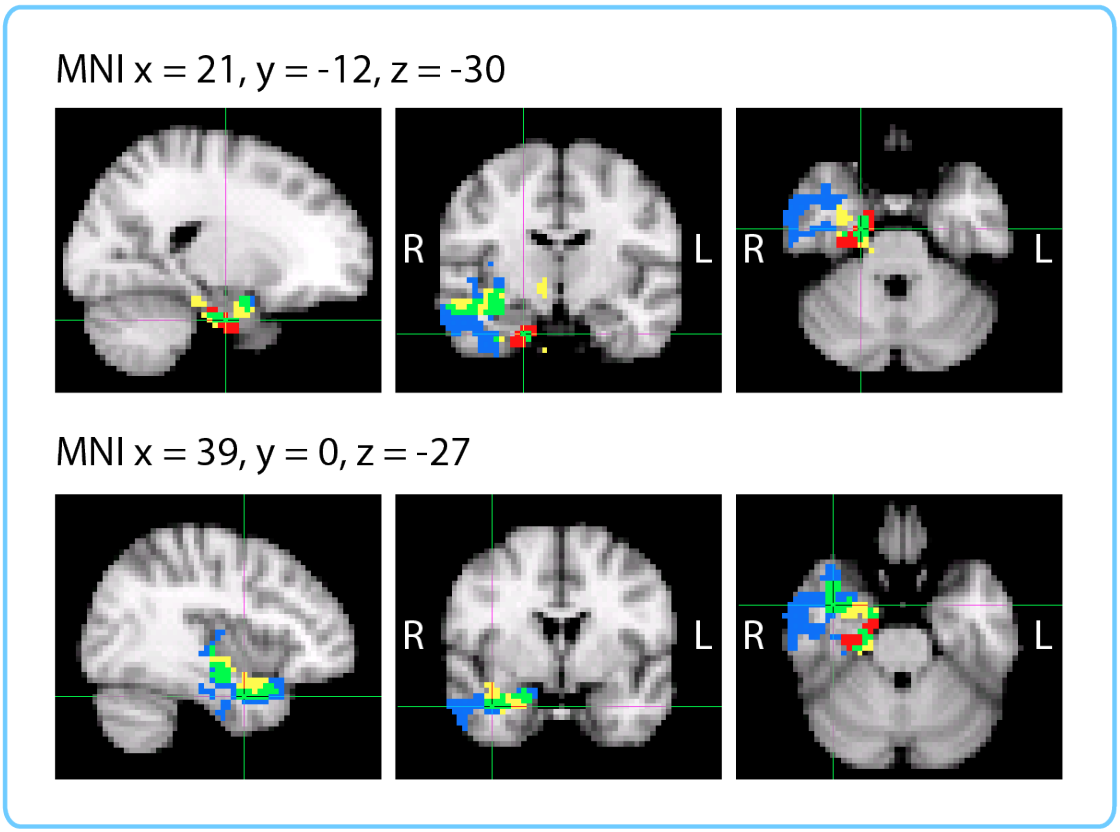


Figure 10 Comparison of ROI and Searchlight results. The within-participant searchlight cluster ($p<0.05$, FWE) is displayed in blue; the within-interval searchlight cluster ($p<0.05$, FWE) is displayed in yellow; voxels that overlap between the searchlights are shown in green. The right entorhinal cortex ($q<0.05$ FDR in both ROI analyses) is displayed in red; voxels that overlap between the within-interval searchlight and the right entorhinal ROI are shown in green.

The difference in the set of regions that passed the significance threshold between the ROI and searchlight analyses is very likely due to the difference in shapes between the searchlight cube and the anatomical masks. Following the anatomy is particularly important for small, elongated regions like entorhinal cortex and caudal ACC, which are unlikely to be perfectly aligned across participants. For the searchlight analyses, the data needed to be transformed to MNI space in order to aggregate the results; consequently, imperfections in alignment can reduce the significance of searchlight results in these regions. On the other hand, anatomical ROI analyses were performed entirely in native space, making them more suitable for idiosyncratically shaped regions.

Patterns of activity in entorhinal cortex change slowly over time

To further probe the idea that the regions we found represent slowly changing contextual features, we assessed whether their patterns of activity change slowly over time relative to the rest of the brain. We focused this analysis on the right entorhinal cortex and left caudal ACC, both of which were significant in the mixed-effects ROI analysis.

We quantified the speed of BOLD signal change in three different ways: 1) a multivariate procedure, 2) a multivariate procedure in which we regressed out ROI size, and 3) a univariate procedure. 1) For the multivariate procedure, we obtained the mean auto-correlation function of the pattern in every region, and took the full-width half-maximum (FWHM) of this function as a measure of how slowly the pattern moves away from itself on average (see *Methods*). 2) Since this analysis was performed on anatomical masks derived from FreeSurfer parcellation,

they varied substantially in size. To ensure that differences in the speed of pattern change were not due to differences in ROI size, we also performed the multivariate procedure after regressing the vector of ROI sizes (number of voxels) out of the vector of FWHM values for each participant. 3) Finally, we performed the above analysis for every voxel individually. Rather than calculating the mean auto-correlation function of the pattern in every region, we calculated the auto-correlation function of every voxel's time course and averaged the auto-correlation functions across all the voxels in a given region. The FWHM was then computed for this mean auto-correlation derived from individual voxel time courses.

Using these three procedures, we compared the FWHMs in the right entorhinal cortex and left caudal ACC with FWHMs in three regions known to be involved in auditory and language processing: the right transverse temporal cortex, which encompasses primary auditory cortex (Destrieux, Fischl, Dale, & Halgren, 2010; Shapleske, Rossell, Woodruff, & David, 1999), the right banks of the superior temporal sulcus and the right superior temporal cortex, which are involved in auditory processing and the early cortical stages of speech perception (Binder et al., 2000; Hickok & Poeppel, 2004).

Table 1 shows the FWHMs in the above regions derived using the three procedures, as well as the ranking of the right entorhinal cortex and left caudal ACC mean FWHMs relative to all the other masks in the brain (84 in total).

Table 1 Speed of pattern change in the right entorhinal cortex and left caudal ACC relative to the rest of the brain. Full-Width Half-Maximum (FWHM) values reflect how slowly patterns of activity (multivariate) or individual voxels (univariate) change over time. The Multivariate (-ROI size) column reflects the slowness of pattern change when controlling for the effect of ROI size.

	Multivariate		Multivariate (-ROI size)		Univariate	
Region	**FWHM (TRs)**	**Ranking**	**FWHM (TRs)**	**Ranking**	**FWHM (TRs)**	**Ranking**
Right entorhinal	*M*=18.9, *SD*=13.8	3rd	*M*=1.2, *SD*=1.9	4th	*M*=23, *SD*=15.6	1st
Left caudal ACC	*M*=8.3, *SD*=1.8	66th	*M*= -0.5, *SD*= 0.5	67th	*M*=9.2, *SD*=3.8	46th
Right transverse temporal cortex	*M*=7.3, *SD*=1.2	80th	*M*= -0.8, *SD*= 0.5	83rd	*M*=7.9, *SD*=1.2	68th
Right banks of superior temporal sulcus	*M*=9.0, *SD*=2.1	48th	*M*= -0.3, *SD*= 0.4	49th	*M*=8.8, *SD*=1.7	61st
Right superior temporal cortex	*M*=11.0, *SD*=3.1	28th	*M*= 0.4, *SD* = 0.6	18th	*M*=10.3, *SD*=2.4	34th

Across all three procedures, a right-tailed Wilcoxon signed-rank test indicated that the FWHMs in the right entorhinal cortex were consistently larger across participants than the FWHMs in the right transverse temporal cortex ($p<0.00005$, $p<0.0005$ and $p<0.00005$), the right banks of the superior temporal sulcus ($p<0.001$, $p<0.001$ and $p<0.0005$) and the right superior temporal cortex ($p<0.005$, $p=0.06$ and $p<0.0005$). Thus, single voxels and multivariate patterns in entorhinal cortex changed consistently more slowly than those in regions involved in auditory and language processing. Moreover, the mean FWHM in the right entorhinal cortex was one of the largest among all 84 regions, ranking 3rd, 4th and 1st in the brain across the three procedures. The other regions with the slowest voxel and pattern change included the temporal pole, medial and lateral orbitofrontal cortex, frontal pole, perirhinal cortex, pars orbitalis and inferior temporal cortex.

On the other hand, the left caudal ACC ranked 66th, 67th and 46th out of 84 regions across the three procedures, suggesting that it did not exhibit slow signal change relative to the rest of the brain. Across the three procedures, the FWHMs in the left caudal ACC were larger than those in the right transverse temporal cortex (p<0.01, p<0.005, and p=0.059), but generally smaller than those in the right banks of the superior temporal sulcus (p=0.97, p=0.96, and p=0.42) and the right superior temporal cortex (p=1.0, p=1.0, p=0.98). Thus, patterns in the left caudal ACC changed only slightly more slowly than those in primary auditory cortex.

Taken together, all three variants of the analysis showed that the right entorhinal cortex, along with other regions of the anterior and medial temporal lobe, orbitofrontal cortex and frontal pole, had the slowest pattern change in the brain. These results do not seem to be due to differences in the sizes of the anatomical masks and suggest that the right anterior MTL regions found most consistently in our ROI and searchlight analyses process information that changes slowly over time. Our findings are consistent with those of Stephens, Honey, & Hasson (2013), who showed that auditory cortex regions processing momentary stimulus features had intrinsically faster dynamics than higher-order regions that integrated information over longer time scales (see also Lerner, Honey, Silbert, & Hasson, 2011).

Story position effects cannot explain the correlation between duration estimates and neural pattern change

We found that duration estimates systematically decreased as a function of position in the story, with earlier intervals being estimated as longer than later intervals (***Figure 11***). The correlation between the estimated duration of an interval and its time in the story was consistently negative across participants (M= -0.40, SD= 0.22; $t(16)$= -7.59, p<0.00001).

This result may be a replication of the positive time-order effect: the finding that people judge earlier durations in a series of durations to be longer than later durations (Block, 1982, 1985; Brown & Stubbs, 1988). The effect has been interpreted to mean that context usually changes more rapidly at the start of a novel episode (Block, 1982, 1986). However, another possibility is that the characteristics of the particular story we picked are driving this result. In our story, there was a strong negative correlation between the mean number of event boundaries per interval and the position of the interval in the story (ρ = -0.77). Thus, the decrease in mean duration estimates with story position may be due to the relationship between the number of event boundaries and duration estimates (see *Behavioral Results*).

If the decrease in duration estimates over time is due to a decrease in the amount of contextual change over the course of the story, we might expect BOLD pattern dissimilarity to decrease over time in the brain regions yielded by our ROI analyses. However, there was no consistent correlation between pattern change during an interval and its time in the story for the right entorhinal cortex (M=0.03, SD=0.21; $t(16)$= 0.65; p=0.53), the right pars orbitalis (M= -0.10, SD=0.22; $t(16)$= -1.83, p=0.09), the left caudal ACC (M= -0.05, SD=0.18; $t(16)$=-1.15,

p=0.27), the right amygdala (*M*= -0.02, *SD*=0.23; *t*(16)= -0.28, *p*=0.78) or the right insula (*M*= -0.08, *SD*=0.25; *t*(16)=-1.34, *p*=0.20). These results suggest that the relationship between duration estimates and pattern dissimilarity in these regions was not driven by a shared effect of story position. Rather, it seems that pattern dissimilarity in these regions correlated with more fine-grained variations in the estimated durations of nearby intervals (***Figure 11***).

To investigate why the above regions did not show the expected decrease in pattern dissimilarity over time, we assessed whether any brain region in the FreeSurfer or MTL atlas might show this effect. There was no brain region whose pattern of activity changed more at the beginning than at the end of the story. Given that we were looking for a slow change in neural signal (unfolding over the entire course of the story), we thought that our high-pass filter might be removing this slow change; to address this possibility, we analyzed the unfiltered data. When we did this, we found that neural pattern change in the unfiltered data showed a consistent correlation in the *opposite* direction: almost all brain patterns changed more at the end of the story than at the beginning, including the CSF and white matter (q<0.05, FDR), suggesting that a signal unrelated to neural processing, such as scanner drift or motion, may cause activity patterns to change more as time passes (see ***Figure 11 – source data 1***). Thus, even if the degree of neural pattern change were decreasing over time, we might not be able to detect this effect, as it would have to overcome a global signal in the opposite direction that is not due to neural activity and that is present everywhere, including the CSF.

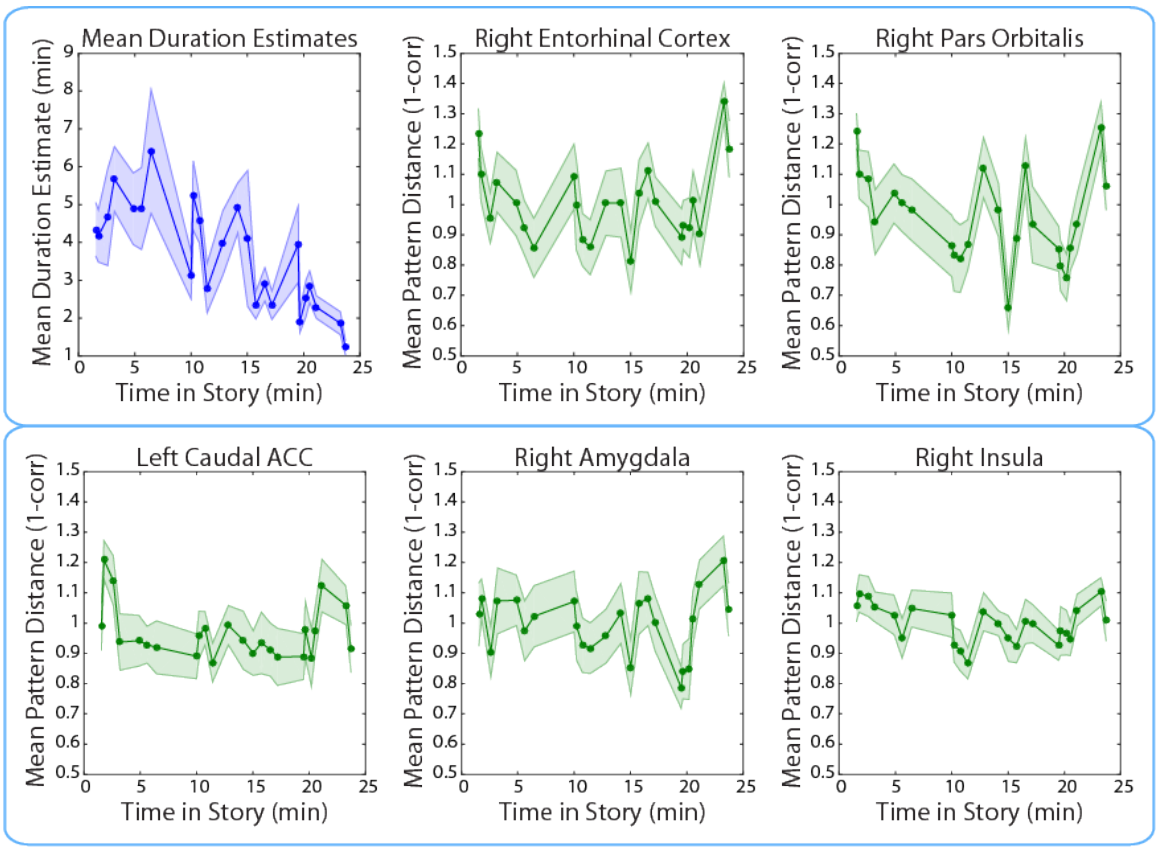


Figure 11 Mean duration estimates and pattern distances (across participants) for all 2-minute intervals as a function of the interval's position in the story. The middle time point of each 2-minute interval (half-way between the two clips delimiting it) was chosen as the x-coordinate.

Replication of Jenkins and Ranganath 2010: activity at encoding predicts accuracy of temporal context memory

As described in the *Methods* ("Time perception test" section), in addition to estimating the elapsed duration between pairs of clips from the story, participants were given an additional test, where they estimated each clip's position on the timeline of the story. The mean correlation (across participants) between the actual and estimated temporal position on the

timeline of the story was r=0.885 (SD = 0.05), suggesting that participants remembered the temporal position of each clip extremely well ($p<10^{-21}$). ***Figure 12*** shows the timeline estimates for a representative participant (top left panel), as well as the absolute residual error associated with each clip (top right panel), group averaged and plotted against time in the story.

This behavioral dataset enabled us reproduce an fMRI analysis from Jenkins and Ranganath (2010), where voxel activity at encoding was correlated with subsequent accuracy in remembering when a trial occurred in the experiment. For each participant, we regressed the estimated timeline position against the actual position and used the absolute value of the residual as a measure of error. We found that the accuracy (negative error) of timeline placements was significantly correlated with encoding activity in large clusters of the left dorsolateral prefrontal cortex and medial prefrontal cortex, including dorsomedial PFC and anterior cingulate (p=0.008, FWE-corrected; Center of Gravity MNI coordinates (x, y, z) in mm: [-20, 34.8, 28.4]; cluster size = 1121 voxels in 3 mm MNI space), as well as sub-threshold clusters in the medial parietal cortex, including precuneus and posterior cingulate (p=0.058, FWE-corrected; Center of Gravity MNI coordinates (x, y, z) in mm: [-10.5, -54, 16.1]; cluster size = 419 voxels), and left superior temporal gyrus (p=0.098, FWE-corrected; Center of Gravity MNI coordinates (x, y, z) in mm: [-56.9, -19.1, -3.72]; cluster size = 270 voxels).

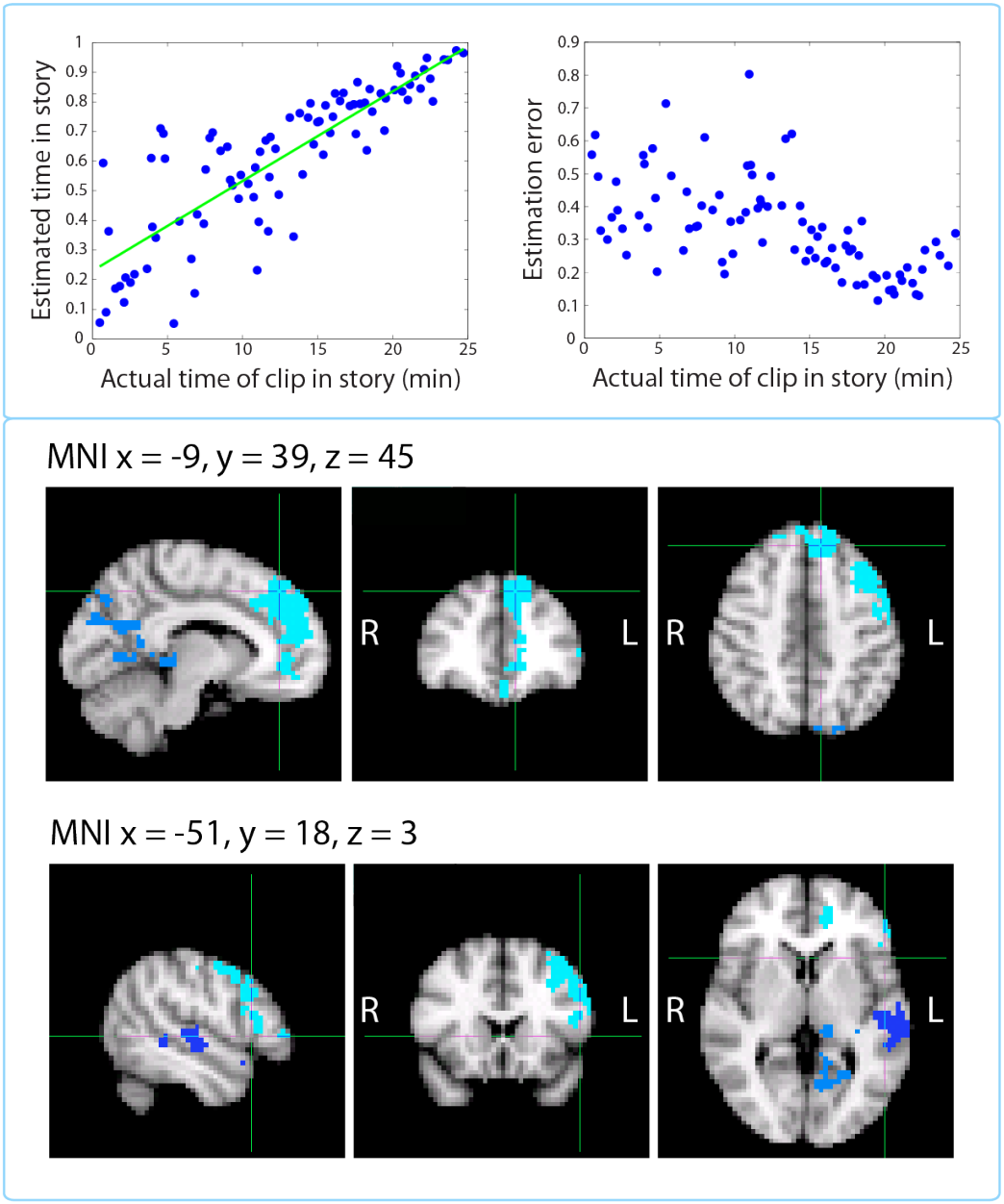


Figure 12 Replication of Jenkins and Ranganath 2010: activity at encoding predicts accuracy of temporal context memory. Top left panel: Timeline estimates for a representative participant. The estimated temporal position of each clip is plotted against its actual position in the story. Top right panel: Group-averaged residual error for each clip plotted against its time in the story. Our behavioral results mimic those of Figure 2 in Jenkins and Ranganath (2010) showing that accuracy increases for clips that occurred later in the story. Bottom panels: Clusters that showed a significant correlation between activity at encoding and subsequent accuracy at placing a clip on the timeline of the story. The prefrontal cluster in light blue was significant (p=0.008, FWE), while the medial parietal cluster (p=0.058, FWE) and the lateral temporal cluster in dark blue (p=0.098, FWE) were trending.

Discussion

While human and animal time perception has been a subject of intense empirical investigation (see Wittmann, 2013), most neuroimaging studies have tested its mechanisms on the scale of milliseconds to seconds and neglected paradigms in which long-term memory plays an important role. Such studies have typically employed *prospective paradigms*, in which participants must deliberately attend to the duration of a stimulus. However, behavioral studies in humans have consistently demonstrated that *retrospective paradigms*, in which participants are asked to estimate the duration of an elapsed interval from memory, tap into different cognitive mechanisms from prospective ones (Hicks et al., 1976; Zakay & Block, 2004; Block & Zakay, 2008). In retrospective paradigms, changes in spatial, emotional and cognitive context tend to modulate estimates of elapsed time (Block, 1992; Block & Reed, 1978; Sahakyan & Smith, 2014; Pollatos et al., 2014).

In the present study, we used changes in patterns of BOLD activity as a proxy for mental context change. We sought to extend previous neuroimaging work by testing whether neural pattern change predicts duration estimates on the scale of several minutes and in a more naturalistic setting, where spatial location, situational inference, characters, and emotional elements can all drive contextual change.

Participants were scanned while they listened to a 25-minute radio story and were subsequently asked how much time (in minutes and seconds) had elapsed between pairs of clips from the story (all pairs were in fact two minutes apart). Using this approach, we were able to probe retrospective duration memory repeatedly within participants without needing to

interrupt the encoding of the story. This allowed us to leverage within-participant variability in neural pattern change and relate it to a participant's retrospective duration estimates.

Using a within-participant anatomical ROI analysis (encompassing 16 regions selected a priori), we found that neural pattern distance in the right entorhinal cortex and right pars orbitalis at the time of encoding was correlated with subsequent duration estimates. Extending this analysis to all anatomical ROIs in cortex revealed an additional effect in the left caudal anterior cingulate cortex (ACC). These results converged qualitatively with the results of our whole-brain searchlight analysis, which revealed a significant cluster spanning the right anterior temporal lobe.

To test our interpretation that duration estimates were driven by contextual change, we asked a separate group of participants to identify event boundaries in the story. We found that the number of event boundaries between two clips was very highly correlated with participants' subsequent duration estimates. Importantly, the number of event boundaries was significantly less correlated with duration estimates for a separate group of "naïve" participants, who had been asked to estimate the elapsed time between clips without first hearing the story. These behavioral experiments provide evidence that retrospective duration estimates were indeed influenced by memory for intervening contextual changes between clips.

In addition, we sought to rule out the possibility that neural pattern distance between two clips reflected only the perceptual or semantic similarity between them, rather than the degree of mental context change. We performed a within-interval analysis, in which pattern distances for the same pair of clips were correlated with duration estimates *across participants*. The

within-interval ROI analysis yielded effects of the same size in the right entorhinal cortex, right amygdala and right insula. The within-interval whole-brain searchlight revealed a significant cluster in the right anterior temporal lobe. Thus, pattern distance in the right anterior temporal lobe, particularly the right entorhinal cortex, predicted variability in duration estimates even when the perceptual and semantic distance of the clips was controlled as much as possible, suggesting that pattern change in these regions may capture idiosyncratic differences in mental context that cannot be predicted from the stimulus alone.

Finally, if neural pattern distance between two clips reflected only the similarity in content between them, rather than abstract contextual similarity, we would expect the correlation between pattern distance and duration estimates to be weakened when controlling for naïve duration estimates, which were based solely on the perceptual and semantic similarity between two clips. Fitting a mixed-effects model to each ROI showed that neural pattern distance in the right entorhinal cortex, along with the left caudal ACC, exhibited a significant effect on duration estimates even when all other factors, including random effects of participants and intervals, as well as naïve duration estimates, were controlled for.

In support of the hypothesis that these regions represent slowly varying contextual information, we found that the right entorhinal cortex, as well as adjacent regions of the MTL, temporal pole and orbitofrontal cortex, had some of the slowest neural pattern change in the entire brain. This is in line with findings that brain regions at the top of the processing hierarchy (furthest from the primary perceptual areas) integrate information over longer time scales and are therefore best suited for representing abstract information extracted from multiple streams of sensory observations (Stephens, Honey, & Hasson, 2013; Lerner et al., 2011).

Our results implicating the right entorhinal cortex in representing context fit well with other results in the literature. Multiple lines of evidence have suggested an important role for the entorhinal cortex in representing relationships between the spatial environment, task and incoming stimuli. Lesions of the lateral entorhinal cortex in rodents have shown that this region is necessary for discriminating between novel and familiar associations of object and place, object and non-spatial context, or place and context, while leaving non-associative forms of memory unaffected (Buckmaster, Eichenbaum, Amaral, Suzuki, & Rapp, 2004; Wilson, Watanabe, Milner, & Ainge, 2013; Wilson, Langston, et al., 2013). Moreover, electrophysiological recordings in rats performing a spatial memory task showed that neurons in the medial entorhinal cortex exhibited greater context sensitivity and greater modulation by task-relevant mnemonic information than hippocampal neurons, while hippocampal neurons carried more specific spatial information (Lipton, White, & Eichenbaum, 2007). Medial entorhinal neurons also exhibited longer firing periods, which led the authors to propose that they could bind a series of hippocampal representations of distinct events (Lipton & Eichenbaum, 2008). Thus, changes in distributed entorhinal activity patterns on the scale of minutes might represent changes in contextual elements that are later retrieved to make duration judgments (for theoretical discussion of the role of entorhinal cortex in contextual representation, see Howard, Fotedar, Datey, & Hasselmo, 2005).

While the right entorhinal cortex was the only medial temporal lobe region that survived FDR correction in both our within-participant and within-interval ROI analyses, our whole-brain searchlights found a significant relationship between pattern change and duration estimates in

two extensive clusters that overlapped in the right hippocampus, the right perirhinal cortex, right amygdala and right temporal pole.

Two previous studies, Noulhiane et al. (2007) and Ezzyat and Davachi (2014), have directly implicated the MTL in retrospective time estimation in humans. Ezzyat and Davachi (2014) scanned participants while they were presented with trial-unique faces and objects on a scene background, which changed every four trials. After each run, participants were asked whether pairs of stimuli had occurred close together or far apart in time (all pairs were about 50 seconds apart). They found that neural pattern distance in the left hippocampus at the time of encoding was greater for pairs of stimuli later rated as "far apart", though only when the stimuli were separated by a scene change. Noulhiane et al. (2007) used a retrospective behavioral paradigm similar to ours in patients with unilateral MTL lesions. In that study, participants were asked to estimate the temporal distance between object pictures that had been randomly inserted into a silent documentary film. They found that the degree of left entorhinal, left perirhinal and left temporopolar cortex damage correlated with the degree to which patients overestimated minutes-long intervals in retrospect. (For related evidence from the animal literature, see Jacobs, Allen, Nguyen, & Fortin, 2013, who showed that bilateral inactivation of the hippocampus impaired rats' ability to discriminate between similarly long durations, such as 8 and 12 minutes, but not between less similar intervals, such as 3 and 12 minutes.)

Our ROI and searchlight results are in line with the above set of findings, and suggest that patients with anterior MTL lesions might be impaired in retrospective time estimation because patterns of activity in entorhinal, perirhinal, and temporopolar cortex encode contextual changes on the scale of minutes. The set of regions we found is more extensive than those in

Ezzyat & Davachi (2014) and mostly right-lateralized. It is possible that the difference in the extent of our effects could be explained by differences in the paradigms that were used. In both the Noulhiane (2007) and Ezzyat & Davachi (2014) studies, the links between objects and their context had to be deliberately constructed. In our study, the clips whose temporal distance participants estimated were excerpts from a story, and therefore strongly linked with a situational, spatial, and emotional context. Thus, it is possible that activity patterns in a more extensive cluster tracked temporal distance estimates because our auditory story caused changes in a broader set of contextual features.

Extending our anatomical ROI analysis to the entire brain showed that pattern change in the left caudal anterior cingulate cortex (ACC) predicted subsequent duration estimates, and this region remained significant in a mixed-effects model controlling for the effect of naïve duration estimates. However, caudal ACC exhibited more rapid pattern change than the anterior and medial temporal lobe, suggesting that it may represent a qualitatively different, faster-changing signal. Caudal ACC activity has been shown to increase in response to shifts in task contingencies (see Shenhav, Botvinick, & Cohen, 2013, for a review) and there is converging evidence that ACC responses are important for adjusting behavior to unexpected changes by increasing attention and learning rate (Bryden, Johnson, Tobia, Kashtelyan, & Roesch, 2011; Behrens, Woolrich, Walton, & Rushworth, 2007; McGuire, Nassar, Gold, & Kable, 2014). Furthermore, O'Reilly et al. (2013) have provided evidence that the ACC only responds to surprising outcomes when they necessitate updating beliefs about the current state of the world. Although the present study was not designed to test such accounts, our findings are consistent with a role for ACC in updating predictive models. Events in the story that prompt

participants to update their beliefs about the characters' situation are also likely to cause changes in cognitive context and therefore overestimation of duration. However, future studies are needed to test this interpretation, for instance by manipulating belief updating independently of surprise and measuring its effect on retrospective duration estimates.

In addition to the anatomical ROI analysis, we performed a whole-brain searchlight that yielded an extensive cluster covering the right anterior temporal lobe, extending from the medial temporal regions described above to the middle temporal gyrus and temporal pole. Prior work has suggested that the middle temporal gyrus and temporal pole are involved in narrative comprehension (Ferstl, Neumann, Bogler, & Von Cramon, 2008; Mar, 2004) and narrative item memory (Hasson, Nusbaum, & Small, 2007; Maguire, Frith, & Morris, 1999). Moreover, Ezzyat and Davachi (2011) found a similarly located cluster (extending from the right perirhinal cortex to the right middle temporal gyrus) to be involved in integrating information within narrative events. In particular, they showed that activity within these regions gradually increases within events and that this increase predicts the degree to which memories become clustered within events. Retrospective time judgments have been shown to increase with the number of events an interval contains (Poynter, 1983; Zakay et al., 1994; Faber & Gennari, 2015), suggesting that brain regions involved in clustering memories by events may carry important information for estimating durations.

Finally, we were able to replicate an analysis by Jenkins & Ranganath (2010), who showed that activity during encoding in the left lateral prefrontal cortex and right anterior hippocampus predicted accuracy in remembering when a trial had occurred in the experiment. Our analysis revealed a cluster in the left dorsolateral prefrontal cortex that is similar to that found in their

study. However, we also found significant clusters in the medial prefrontal and medial parietal cortex that are part of the Default Mode Network. These regions may be important for maintaining narrative information over minutes-long timescales (Lerner et al., 2011; Hasson, Chen, & Honey, 2015; Chen et al., 2015), which might explain why their activity predicted temporal context memory for clips from an auditory story, but did not appear in Jenkins & Ranganath (2010), where participants recalled the timing of trials which were not linked by a narrative. Moreover, our clusters overlap highly with the "posterior medial network" (Ritchey & Ranganath, 2012), which has been consistently implicated in episodic memory, episodic simulation and theory of mind.

Conclusion

After probing human participants' time perception for intervals from an auditory story they had just heard, we found substantial variability in subjective estimates of the passage of time. This variability was significantly correlated with changes in BOLD activity patterns in the right anterior temporal lobe, particularly the right entorhinal cortex, between the start and end of each interval. Control experiments demonstrated that duration estimates were strongly driven by contextual boundaries and that the relationship between neural distance and behavior still held when we controlled for the perceptual/semantic similarity of the clips. Our findings suggest that patterns of activity in these regions might encode contextual information that participants can later retrieve to infer the durations of intervals on the scale of minutes. Additional work is needed to assess how these regions contribute to representing particular contextual features (such as physical environment, abstract task states, and emotional states) and whether changes in each of these features affect retrospective duration estimates differently.

Methods

Participants

18 participants (13 female) took part in the study. All participants were recruited from the Princeton undergraduate and graduate student population and were between 18 and 31 years of age (mean = 22 years). All participants were screened to ensure no neurological or psychiatric disorders. Written informed consent was obtained for all participants in accordance with the Princeton Institutional Review Board regulations. Participants were compensated $20/hour for the scanning session, and $12/hour for the behavioral session.

Given that no previous studies had related neural pattern change during a naturalistic stimulus to subsequent duration estimates for minutes-long intervals, we could not a priori estimate the variance in the pattern change signal, the variance in duration estimates, or the correlation between them. Therefore, rather than performing a power analysis, we chose a sample size that was in the same range as previous fMRI studies that had used naturalistic stimuli to study memory (Lerner et al., 2011, *n*=11 per condition; Chen et al., 2015, *n*=13, 14 and 24 per condition; Chen, Leong, Norman, & Hasson, 2016, *n*=22 (5 excluded)), as well as fMRI studies that had related neural pattern distance to mnemonic judgments (Ezzyat & Davachi, 2011, *n*=19; Jenkins & Ranganath, 2010, *n*=16 (1 excluded); Ezzyat & Davachi, 2014, *n*=21 (3 excluded), Jenkins & Ranganath, 2016, *n*=17).

Experimental Design and Stimuli

The experiment consisted of two parts: an approximately 40-minute session in the MRI scanner, during which participants listened to the auditory story, followed immediately by a 1-hour behavioral session, during which participants completed a time perception test on the story they had just heard. ***Figure 1*** illustrates the experimental procedure.

fMRI session

Prior to the fMRI session, participants were instructed to listen carefully to the auditory story while in the scanner, because they might be asked questions about it later. The nature of the follow-up questions was unknown to the participants. While in the scanner, participants listened to a 25-minute-long radio adaptation of a science fiction story called "Tunnel Under the World" (written by Frederik Pohl), originally aired on the radio drama series, "X Minus One", in 1956.

Time perception test

After leaving the scanner, participants were surprised with a time perception test, presented on a laptop with the Psychophysics toolbox for Matlab (Brainard, 1997; Pelli, 1997). For each of 43 questions, participants listened to a 10 s clip from the story, followed by another 10 s clip, and were asked to estimate how much time had passed between the first and second clips when they initially heard the story. Participants were specifically asked to estimate how much time had passed in their own lives, rather than how much narrative time had passed in the story. They were also asked to make the judgments as intuitively as possible, without

resorting to deductive reasoning about the sequence of events that unfolded in between the two excerpts.

Participants had complete control over the pacing of the test. On each question, they initiated the playing of the clips, and were able to replay the clips if they missed them the first time. They could take as long as they wished to enter their duration estimates (in minutes and seconds), using the keyboard. Clip pairs were identical across participants, but the order in which the pairs were presented was randomized.

To control for the objective passage of time, we ensured that 24 of the clip pairs were 2 minutes apart and 19 of the pairs were 6 minutes apart. Debriefing showed that participants were unaware of this manipulation, and the high variability of duration estimates for both the 2 and 6-minute intervals further confirmed that they were unaware of the fixed interval durations.

After participants had provided duration estimates for all 43 intervals, the 86 clips that had delimited those intervals were replayed in a random order (unpaired), and participants were asked to place each clip on the timeline of the story. For each of the 86 questions, a white line appeared on a black background, representing the full length of the story. Participants could place their cursor at any point on that line, followed by the Enter key. After each placement, they were asked to provide a confidence rating on a scale of 1 to 5, reflecting their confidence about that clip's place in the story. Participants were instructed to base the confidence rating on their certainty of when that clip occurred in the story, rather than on the vividness of the memory for that clip. While the exact placement of each clip on the timeline was not used in

the fMRI analysis, confidence ratings were used to exclude clips whose temporal context participants had forgotten.

Please note: the first of our 18 participants completed a version of the time perception test that differed only in the following way: the specific intervals in the story whose duration was asked about were different. In all other respects (half of the intervals were 2 minutes while the other half were 6 minutes apart), the behavioral test was identical to the subsequent 17 participants. For this reason, however, any analyses where duration estimates are compared across participants were performed on 17 rather than 18 participants. Any within-participant analyses were performed on all 18 data sets.

Naïve time perception test

To address the concern that participants were estimating temporal distance between two clips based purely on the content of the clips (rather than their memory of when the clips had occurred in the story), we administered an identical time perception test to a separate group of 17 participants who had never heard the story. Naïve participants were asked to try their best to guess how much time passed between each pair of clips during the original telling of the story, even though they had never heard the story. Participants were told the length of the story (25 minutes, 33 seconds) and informed that the maximum distance between two clips could not exceed that duration.

Event boundary test

A separate group of 9 participants were asked to listen to the same story and to press the space bar every time they thought an event had ended and a new event was beginning. This test was purely behavioral and fMRI data were not collected for these participants.

Behavioral Data Analysis

Significance of correlation between duration estimates and event boundaries

To assess whether the number of event boundaries in an interval predicted duration estimates for that interval, we related our original participants' duration estimates with event boundary data collected from a separate group of 9 participants. For each 2-minute interval from the time perception test, we counted the number of event boundaries that a participant had indicated during that interval and averaged that number across the 9 participants. This resulted in a mean number of event boundaries per interval, which was then correlated with the mean estimated duration of that interval from our original participants.

To assess the statistical significance of this correlation, we performed a bootstrapping procedure on the duration estimates. We obtained 1000 bootstrap samples, each time selecting with replacement a different subset of n individuals from our pool of n participants. The duration estimates for each subset were averaged across participants and correlated with the mean number of event boundaries. The upper limit *(ul*) for an x% confidence interval was set to the value of the Pearson correlation in percentile x% of the bootstrap distribution; the lower limit *(ll*) for the confidence interval was set to the value of the beta score in percentile

100-*x* of this distribution. Confidence intervals that did not encompass zero were considered reliable at the given level of confidence.

Significance of difference in correlations with event boundaries between original duration estimates and naïve duration estimates

We hypothesized that duration estimates from our original participants (who had actually heard the story) would be significantly more correlated with the number of event boundaries between two clips than duration estimates from our naïve participants, who had never heard the story. To assess the significance of the difference in correlations, we computed the r_{diff} (empirical difference), as well as the upper confidence limits (ul_{diff}) and lower confidence limits (ll_{diff}) for the difference between the two correlations. We used the following formulae (Zou, 2007; Poppenk & Norman, 2012) for two bootstrapped correlation confidence intervals:

$$r_{diff} = r_1 - r_2$$

$$ll_{diff} = r_1 - r_2 - \sqrt{(r_1 - ll_1)^2 + (ul_2 - r_2)^2}$$

$$ul_{diff} = r_1 - r_2 + \sqrt{(ul_1 - r_1)^2 + (r_2 - ll_2)^2}$$

The upper (ul_1, ul_2) and lower limits (ll_1, ll_2) for a 95% confidence interval of each group's correlation were calculated as described above.

Reliability of duration estimates across participants within and between groups

We hypothesized that both our original participants and the naïve participants (who had never heard the story) would use consistent strategies to estimate the temporal distance between two clips, but that these strategies would differ across groups. If this is the case, duration estimates should be more reliable across participants within groups than across participants between groups.

To assess within-group reliability, we correlated each participant's duration estimates with the mean of the other participants' estimates. These correlations were then averaged across participants within a group to obtain a mean within-group ISC (inter-subject correlation). The between-group reliability was calculated by correlating each participant's duration estimates from one group (e.g., the original participants) with the mean duration estimates from the other group (e.g., the naïve participants). These correlations were then also averaged across participants to obtain a mean between-group ISC. Confidence intervals for the mean between-group ISC were calculated by bootstrapping the duration estimates from both groups 10,000 times, each time selecting with replacement a different subset of *n* individuals from our pool of *n* participants. The between-group ISCs were calculated for each bootstrap sample and averaged across participants, resulting in a distribution of 10,000 mean between-group ISCs. Confidence intervals for the within-group ISC were obtained in a similar manner.

To assess the significance of the difference between the mean within-group ISC and the mean between-group ISC, we compared the empirical difference with a null distribution of differences. Group labels (naïve participants vs. original participants) were scrambled 10,000

times, such that each participant's duration estimates were randomly assigned to either the naïve group or to the original group. The difference between the mean within-group ISC and the mean between-group ISC was then computed for these two random groups. Using this null distribution of ISC differences, we calculated a p-value based on the number of permutations that yielded a greater difference than the empirical difference.

Please note that the within-group and between-group correlations could be compared only because the group sizes were identical (17 participants in each) and because the within-group correlations were equally strong for the original and naïve groups (M=0.43, SD=0.25, 95% CI=[0.37, 0.58] vs. M=0.43, SD=0.18, 95% CI [0.40, 0.56]). Since the within-group ISCs are comparable, we can infer that the significant difference between the within-group and between-group reliability reflects a difference in the signals (strategies) underlying the two groups of duration estimates (Chow, Chen, & Hasson, 2015), rather than a difference in within-group reliability.

MRI Acquisition

Participants were scanned in a 3T full-body MRI scanner (Skyra, Siemens) with a 20-channel head coil. Functional images were acquired using a T2*-weighted echo planer imaging (EPI) pulse sequence (repetition time [TR], 1500 ms; echo time [TE], 28 ms; flip angle, 64°), each volume comprising 27 slices of 4 mm thickness. In-plane resolution was 3×3 mm^2 (field of view [FOV], 192×192 mm^2). Slice acquisition order was interleaved. Anatomical images were acquired using a T1-weighted magnetization-prepared rapid-acquisition gradient echo

(MPRAGE) pulse sequence (TR, 2300 ms; TE, 3.08 ms; flip angle 9°; 0.89 mm^3 resolution; FOV, 256 mm^2). Participants' heads were stabilized with foam padding to minimize head movement. Auditory stimuli were presented using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997). Participants were provided with MRI compatible in-ear mono earbuds (Sensimetrics Model S14), which provided the same audio input to each ear. MRI-safe passive noise-canceling headphones were placed over the earbuds for additional protection against noise.

fMRI Data Preprocessing

FMRI data processing was carried out using FEAT (FMRI Expert Analysis Tool) Version 5.98, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). The following procedure was applied: motion correction using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002); slice-timing correction using Fourier-space time-series phase-shifting; non-brain removal using BET (Smith, 2002); spatial smoothing using a Gaussian kernel of FWHM 6.0 mm; grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor; and high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=240.0 s). The procedure for selecting the high-pass filter is described below. Preprocessed data were kept in the native functional space for all analyses, except for the within-interval searchlight analysis, which was performed across participants.

Preprocessed data were then despiked using the following procedure: for each voxel, data points that deviated from the mean by more than 5 times the inter-quartile range were removed and replaced using cubic interpolation.

Procedure for obtaining anatomical masks: FreeSurfer and MTL segmentation

Segmentation was performed in a semi-automated fashion using the FreeSurfer image analysis suite, which is documented and available online (version 5.1; http://surfer.nmr.mgh.harvard.edu) with details described previously (e.g. Fischl et al., 2004; Poppenk & Norman, 2014). Briefly, this processing includes removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Ségonne et al., 2004), automated Talairach transformation, intensity normalization (Sled, Zijdenbos, & Evans, 1998), tessellation of the grey matter / white matter boundary, automated topology correction (Fischl, Liu, & Dale, 2001; Segonne, Pacheco, & Fischl, 2007), surface deformation following intensity gradients (Fischl & Dale, 2000), parcellation of cortex into units based on gyral and sulcal structure (Desikan et al., 2006; Fischl et al., 2004), and creation of a variety of surface-based data, including maps of curvature and sulcal depth.

We resampled and aligned FreeSurfer segmentations of all grey matter, white matter, and cerebrospinal fluid (CSF) regions to native functional image space for use as anatomical masks. Anatomical regions were segmented according to the Desikan-Killiany Atlas (Desikan et al., 2006).

It is important to note that the medial temporal lobe (MTL) masks in the Desikan-Killiany Atlas do not match the canonical anatomical distinctions in the literature. For example, the parahippocampal gyrus mask comprises the medial part of the parahippocampal cortex and the posterior part of the entorhinal cortex. Therefore, instead of the FreeSurfer MTL masks, we

used a probabilistic MTL atlas developed by Hindy & Turk-Browne (2015). MTL regions, including perirhinal cortex, entorhinal cortex and parahippocampal cortex were defined probabilistically in MNI space, based on a database of manual MTL segmentations from a separate set of 24 participants. Manual segmentations were created on T_2-weighted turbo spin-echo images using anatomical landmarks (Duvernoy, 2005; Carr, Rissman, & Wagner, 2010; Schapiro, Kustner, & Turk-Browne, 2012) and then registered to an MNI template. Finally, nonlinear registration (FNIRT; Andersson, Jenkinson, & Smith, 2007) was used to register the masks from MNI space to each participant's native space. After registration, voxels with a probability greater than 0.3 of being in a region were assigned to that ROI.

Residualization of non-neuronal signal sources

Slow changes of respiration over time (RV) have been shown to induce robust changes in the BOLD signal (Chang, Cunningham, & Glover, 2009) in many areas around the cerebral midline. To minimize signal change unrelated to neural activity, we used multiple linear regression to project out 3 nuisance variables from the BOLD data (Behzadi, Restom, Liau, & Liu, 2007; Silbert, Honey, Simony, Poeppel, & Hasson, 2014). Nuisance regressors were:

1) the average time course of high standard deviation voxels (voxels with the top 1% largest standard deviation), as these voxels tend to have the highest fractional variance of physiological noise (e.g., cardiac and respiratory components) and are likely near blood vessels (Behzadi et al., 2007),

2) the average BOLD signal measured in CSF,

3) the average white matter signal.

All masks (grey matter, white matter and CSF) were obtained from the FreeSurfer segmentation procedure described above. The beneficial effects of this residualization procedure on the signal-to-noise ratio are shown in ***Figure 13***. Note that this procedure was always applied after removal of low-frequency components using the high-pass filter (see below).

Methodological challenges with analyzing pattern distance over long time scales: Selection of temporal high-pass filter cut-off

Because we were interested in the aspect of neural activity that changes slowly over time (reflecting gradual changes in context), we could not use a standard high-pass filter (with a cut-off period on the order of 120 s), as it would remove components of the signal that evolve on the scale of minutes. Thus, we were faced with the challenge of preserving slower components of the BOLD signal that reflect neural activity, while removing low-frequency components attributable to non-neuronal noise, including scanner drift and physiological noise (such as low-frequency respiratory variation and heart rate variation). Physiological noise (and a substantial component of scanner noise) was factored out using the residualization procedure described above. This enabled us to select a gentler high-pass filter than is generally used in the literature.

We then performed a separate analysis to determine the optimal high-pass filter cut-off period, i.e. the lowest frequency cut-off that still enabled us to remove most of the non-neuronal noise. This analysis relies on the idea that, when participants listen to the same story or watch the same film, the signal in brain regions processing the story is highly correlated

across participants (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004). While such correlations should not be present in CSF or white matter, spurious inter-subject correlations in these regions can arise due to low-frequency noise. In addition, listening to the same story could induce correlated motion across participants, but these correlations would also be present in CSF and white matter. Thus, we searched for a high-pass filter that could remove nonspecific correlations in CSF and white matter, while preserving correlations in brain regions known to be important for processing the stimulus. For each participant, the inter-subject correlation (ISC) of a brain region was defined as the correlation between that participant's ROI time course (averaged over voxels in that region) with the average time course of all the other participants (Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Lerner et al., 2011).

Since the functional scan length was 1560 s (26 minutes), high-pass filter cut-off periods of 140 s, 240 s, 300 s, 400 s, 480 s, 600 s and 720 s were attempted. The minimal cut-off attempted, 140 s, was the cut-off used in several previous studies with naturalistic stimuli (e.g. Lerner et al., 2011), while 720 s represented approximately half of the scan duration and was the longest cut-off that could reasonably make a difference to data quality.

Given that roughly half the clip pairs in our time perception test were 2 minutes apart and the other half were 6 minutes apart, we hoped to find a filter that would allow us to measure pattern distances at both of these time scales. However, we were unable to find a high-pass filter that would allow us to examine activity patterns that were 6 minutes (360 s) apart. In order to meaningfully measure distances between neural patterns that are 360 s apart, the Nyquist theorem suggests we would need a high-pass filter cut-off of 720 s or larger. However, plotting ISC as a function of high-pass filter (***Figure 13***) showed that a cut-off like 720 s was not

able to remove inter-subject correlations in the CSF, which remained of the same magnitude as those in some grey matter regions. We concluded that pattern distances at the 6-minute time scale are too confounded with low-frequency noise (as reflected in spurious correlations in the CSF), and therefore restricted our analysis to intervals that were 2 minutes long.

According to the Nyquist theorem, we need a filter cut-off of 4 minutes (240 s) or longer in order to measure distances between patterns that are 2 minutes apart (120 s). Out of the filters tested (240 s – 720 s), a cut-off of 480 s was selected to be the gentlest (i.e. the longest) filter that reduced the magnitude of inter-subject correlations in ventricles and CSF, such that they were significantly below the correlations in most grey matter regions.

Figure 13 illustrates that, even for regions like the hippocampus – with relatively low inter-subject correlations – the 480 s filter cut-off, combined with the residualization procedure, succeeded at raising the grey matter ISCs significantly above those of the white matter and CSF.

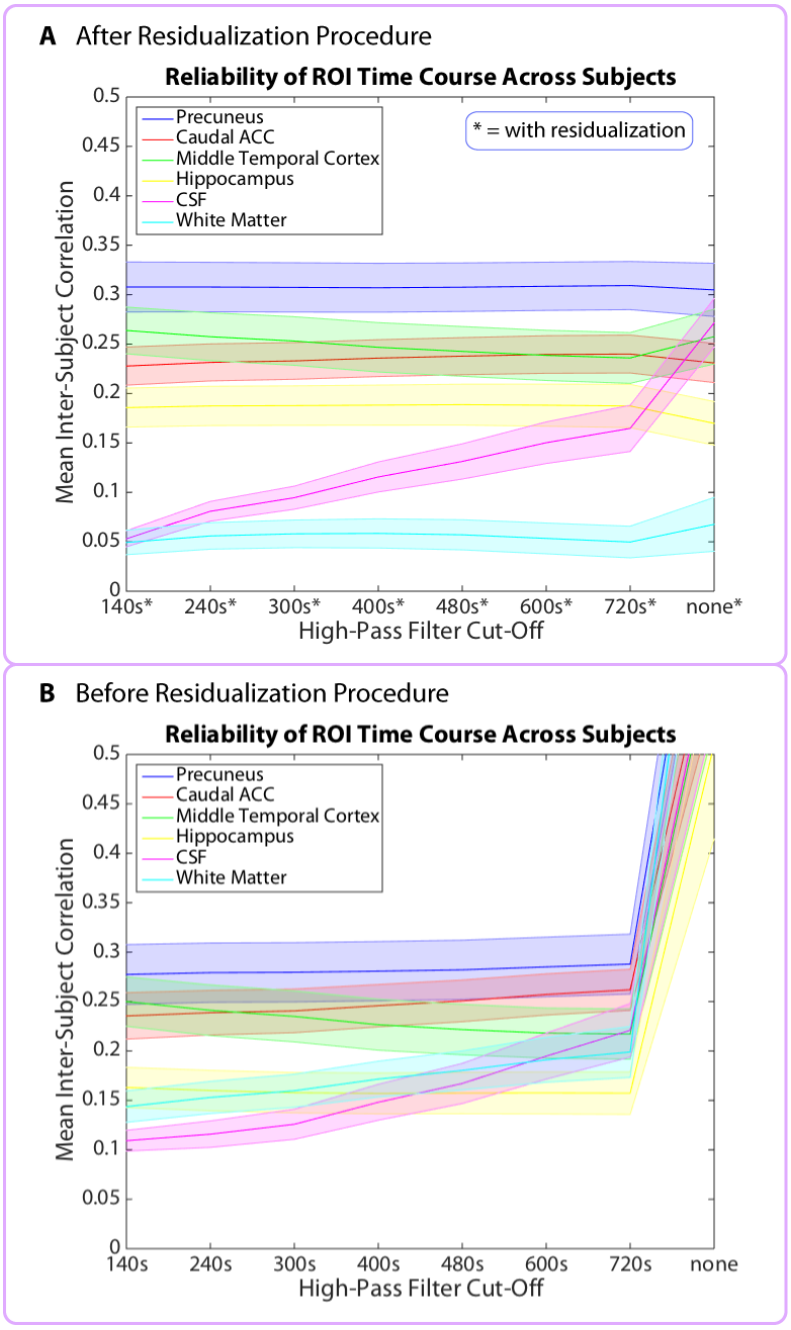


Figure 13 Mean inter-subject correlations (ISCs) for 6 representative brain regions as a function of the high-pass filter cut-off. Shaded error bars represent standard errors of the mean (across participants). Top panel A shows the mean ISCs after the residualization procedure has been applied (see "Residualization of non-neuronal signal sources"). The 480 s cut-off was the gentlest filter for which all of the grey matter regions listed above showed ISC values significantly above those in the CSF. Bottom panel B shows the mean ISCs prior to the residualization procedure. Without residualization, the ISCs of some grey matter regions never rise significantly above those in the white matter and CSF. Note that without high-pass filtering ("none") or residualization, all brain regions displayed spuriously high ISCs.

fMRI Data Analysis

Within-participant correlation between pattern change and duration estimates

Our primary hypothesis was that greater pattern dissimilarity between two clips (at the time of encoding) would correlate with greater subsequent duration estimates. For each pair of clips from the time perception test, we located the TRs (volumes) corresponding to when the participant first heard those clips and extracted the activity patterns for each ROI at those time points. Since the auditory clips were between 5 s and 10 s in duration (corresponding to about 5 volumes), we averaged the patterns over 5 consecutive TRs for every clip, with the 5-TR window centered on the middle of each clip.

We then related the pattern distance between the two clips at encoding to how much time the participant thought passed between them. Specifically, we calculated the dissimilarity (1 – Pearson correlation) between the two averaged activity patterns. The pattern dissimilarity scores for a given region were then correlated with that participant's subsequent duration estimates. This was performed separately for every ROI and searchlight (***Figure 4***). We thus obtained a Pearson correlation score for every ROI in every participant. All Pearson correlation coefficients were Fisher-transformed prior to statistical testing (Fisher, 1915).

To assess the reliability of the correlation across participants for a given ROI, we ran a phase-randomization procedure, which is described in detail below. The results of the phase-randomization procedure were then subjected to multiple comparisons correction.

Removing low-confidence intervals

If a participant could not remember when in the story a particular clip had occurred, it would be difficult for them to estimate the temporal distance between that clip and another clip. It is possible that participants would invoke different retrieval strategies in such cases (for instance, they might base their duration estimates purely on the content of the clips, without recollecting their context). It is also possible that such estimates could be random guesses. To filter out guesses, we used the confidence ratings collected after the time perception test, in which participants rated how well they could remember when in the story each individual clip had occurred. Specifically, we located the participant's confidence for the two clips delimiting each temporal interval, and took the smaller of the two ratings as the confidence for that interval. We performed the main analysis relating neural drift to time estimation (described below) only on high-confidence intervals, removing pairs of clips with the lowest confidence. Since participants calibrated their confidence ratings differently (some were more prone to rate their confidence as 4/5, while others were more prone to rate it as 2/5), we picked the confidence threshold for each participant that removed at least 33% of the intervals with the lowest confidence, while preserving at least 33% of the intervals with the highest confidence. Our behavioral analysis (see *Results*) shows that participants' duration estimates were significantly more accurate for high-confidence intervals than when all intervals were included.

Statistical analysis of correlations between pattern change and behavior

Because of the presence of long-range temporal autocorrelation in the BOLD signal (Zarahn et al., 1997), the statistical likelihood of each observed correlation (between neural distance and duration estimates) was assessed using a permutation procedure based on surrogate data. The surrogate data were generated using phase randomization (Theiler et al., 1992). Phase-randomized surrogates have the same autocorrelation as the original signal.

Since our analysis measures pattern change over multiple voxels, rather than the time course of a single voxel, we generated surrogate time courses of pattern change (***Figure 4 – Supplement 1*** shows how that time course was obtained). Having extracted the time course of pattern change for each ROI, we applied a Fourier transform to that signal. To randomize its phases, we multiplied each complex amplitude by $e^{j\phi}$, where ϕ is independently chosen for each frequency from the interval [0, 2π]. In order for the inverse Fourier transform to be real (no imaginary components), we symmetrized the phases, so that $\phi(f) = -\phi(-f)$. Finally, we took the inverse Fourier transform to produce the surrogate time courses.

Each surrogate dataset was analyzed in the same manner as the empirical data: pattern dissimilarity between each pair of clips was correlated with duration estimates. Thus, we generated a distribution of 10,000 null correlations for every ROI in every participant (see ***Figure 4 – Supplement 1***). As above, all correlation coefficients were Fisher-transformed to ensure that they follow a Gaussian distribution. For every ROI, we were then able to compare the empirical Pearson correlation with the distribution of null correlations. We calculated a Z-value for every participant:

$$Z - value = \frac{empirical\ correlation - mean(null\ correlations)}{standard\ deviation(null\ correlations)}$$

A large positive Z-value implies that the empirical correlation is large relative to the distribution of null correlations. To assess whether the Z-values for a given ROI were reliably positive across participants, we performed a right-tailed t-test against 0. The p-values from the above t-test were then subjected to multiple comparisons correction. For anatomical ROIs (derived from the FreeSurfer and MTL atlases), we used MATLAB's fdr_bky.m function, which executes the "two-stage" Benjamini, Krieger, & Yekutieli (2006) procedure for controlling the false discovery rate (FDR) of a family of hypothesis tests. The procedure implemented by this function is more powerful than the original Benjamini & Hochberg (1995) procedure when a considerable percentage of the hypotheses in the family are false. For the searchlight analysis, we controlled the family-wise error (FWE) rate, as described below.

ROI selection

The literature reviewed above suggests that the MTL, lateral prefrontal cortex, insula, putamen and inferior parietal cortex might all process information important for inferring the duration of past events. We therefore performed an ROI analysis on the following regions, derived from both the FreeSurfer and MTL atlases: hippocampus, parahippocampal cortex, entorhinal cortex, perirhinal cortex, amygdala, superior frontal cortex, caudal and rostral middle frontal gyrus (dorsolateral prefrontal cortex), pars opercularis (frontal operculum), pars triangularis, pars orbitalis, lateral orbitofrontal cortex, frontal pole, insula, putamen and inferior

parietal cortex. This resulted in an analysis on 16 regions of interest (in each hemisphere) motivated by the literature. ROIs with q-values < 0.05 (FDR) are reported as significant.

As part of an exploratory, whole-brain search, we also ran the same analysis on all grey matter regions in the Desikan-Killiany Atlas, which contained 42 regions in each hemisphere, including the ones mentioned above (see *FreeSurfer Segmentation and MTL Segmentation*). The complete list of regions can be found in ***Figure 5 – source data 1***. For the exploratory analysis, we report regions with q-values < 0.1 (FDR).

Within-interval correlation between pattern change and duration estimates

Our main analysis verified whether the pattern distance between two clips was correlated with duration estimates in a given participant and then aggregated the results across participants. To address the concern that pattern distance between two clips might reflect only the difference in story content between those clips (rather than change in abstract factors like mental context), we performed the same analysis for a given interval across participants and aggregated the results across intervals. Since this analysis is performed within intervals, it ensures that story content is held constant across participants, such that differences in pattern distances and duration estimates are due to individual differences only. To ensure that pattern distances and duration estimates were comparable across participants, all vectors were z-scored within participants. The Pearson correlation between pattern distances and duration estimates across participants was then calculated for every 2-minute interval in every ROI.

As for the within-participant analysis, this procedure was performed on high-confidence intervals. For each interval, we only included participants who had confidently recollected the temporal position of the two clips delimiting that particular interval.

The significance of each correlation score was assessed using a permutation test: 10,000 null correlations were obtained by scrambling the duration estimates across participants, such that a given participant's duration estimate was matched with a different participant's pattern distance. (Since this analysis was performed across participants, it was not necessary to generate phase-randomized pattern distance time courses – the auto-correlation in the BOLD signal for a given participant only represents a concern for the within-participant analysis.)

As above, a Z-value was obtained for every interval, reflecting the degree to which the empirical correlation was higher than the distribution of null correlations. Finally, a right-tailed t-test was performed to assess whether a given ROI's Z-values were reliably positive across intervals. The p-values from this t-test were subjected to multiple comparisons correction using FDR.

To compare effect sizes between the within-interval and within-participants analyses, we calculated Cohen's *d* for a region as:

$$Cohen's\ d = \frac{Mean\ \rho\ (across\ participants\ or\ intervals)}{Standard\ deviation\ of\ \rho}$$

where ρ is the Pearson's correlation between pattern distance and duration estimates. (Using the Z-values derived from the permutation procedures rather than the raw correlation coefficients yielded practically identical results.)

Mixed-Effects Model Accounting for Naïve Duration Estimates

We analyzed our data using a hierarchical linear regression model (Gelman & Hill, 2006). Known in different fields as hierarchical, mixed, or multi-level models, such regressions correctly account for non-independence of repeated observations of the same subject and stimulus (in our case, interval). In doing this they estimate the population effects (coefficients) of interest even assuming that individual subjects or items (henceforth, collectively "groups") may have idiosyncratic perturbations from the population, and that those perturbations may be correlated within a group. They are a generalization of approaches that treat all observations as independent (e.g. t-test, ANOVA, linear regression), as well as of approaches that can take into account the non-independence across a single grouping factor (e.g. repeated-measures ANOVA), and are more conservative than any of the above (Barr, Levy, Scheepers, & Tily, 2013)[2].

Formally, the model is the following:

$$y_i = X_i(\beta + s_{j[i]} + m_{k[i]}) + \epsilon$$

$$s_j \sim N(0, \Sigma_S), \qquad m_k \sim N(0, \Sigma_M), \qquad \epsilon \sim N(0, \sigma)$$

Here, y_i is the *i*th observed duration judgment, X_i is a matrix of predictors (neural pattern distance) and covariates (naïve duration estimates), β_i is a vector of coefficients (as in conventional linear regression), $j[i]$ is the subject of the *i*th observation, so that $s_{j[i]}$ is a

[2] More precisely, methods that do assume observation independence are anti-conservative in the presence of correlated observations.

subject-specific perturbation of all of the coefficients, and $m_{k[i]}$ is similarly a vector of item-specific perturbations.

This model is undefined when either the subject or item effects approach zero (either because there is truly no variability, or more realistically when there is insufficient data to estimate this variability). Since such rich models often fail to converge or approach singularity given typical psychological datasets (Bates, Kliegl, Vasishth, & Baayen, 2015), we imposed a weak Wishart prior on the group covariances, which regularizes the model away from singularity (Chung, Gelman, Rabe-Hesketh, Liu, & Dorie, 2015). This weak, boundary-avoiding prior on our random effects covariance structure regularizes the model towards simpler random effects structures unless the data suggests otherwise (Chung et al., 2015). All models converged under this prior. This fitting procedure was implemented using the R package `blme` (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013), which extends the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) and performs maximum-a-posteriori estimation of linear mixed-effects models.

Please note that we also verified that our results were replicable using an alternative fitting procedure suggested by Bates, Kliegl, et al. (2015). We used the `lme4` package to fit the 'maximal' model (in the sense of Barr et al., 2013) and removed zero-variance random effects terms until the model converged and until the estimated random effects covariance matrix was full-rank, indicating a non-degenerate estimate. We obtained highly consistent results using both fitting procedures. In the *Results* section, we report only the first procedure, which has been found to be more conservative (Chung et al., 2015). Chung et al. (2015) report: "Uncertainty for the fixed coefficients is less underestimated than under classical ML or

restricted maximum likelihood estimation." Indeed, our effects were very slightly stronger using the second procedure (Bates, Kliegl, et al., 2015). Both sets of results can be found in ***Figure 7 – source data 1***.

Finally, the duration estimates are bounded at zero and positively skewed, which resulted in heteroskedastic residuals. To mitigate this, we power-transformed the duration estimates using the Box-Cox power transformation (Box & Cox, 1964). We picked the exponent λ for each model by maximizing the profile likelihood in a model without group effects (though see e.g. Gurka, Edwards, Muller, & Kupper (2006) for an extension to the hierarchical case).

In R formula notation, a model of the following form was fit to the data from each region of interest:

$$Transformed\ Duration\ estimates \sim 1 + NaiveEstimates + NeuralPatternDistance + (1 + NaiveEstimates + NeuralPatternDistance \mid Subject) + (1 + NeuralPatternDistance \mid Interval)$$

Please note that participants from the original experiment could not be "matched" with participants from the naïve experiment. For this reason, naïve duration estimates were group-averaged and the mean vector of naïve estimates was placed as a covariate in the model. The above formula shows that the slope of the relationship between naïve estimates and original duration estimates was allowed to vary by subject (i.e. each participant's duration estimates might be differently related to naïve group mean). On the other hand, the slope for naïve estimates could not vary by interval, since naïve estimates did not vary by subject.

We computed 0.95 confidence intervals of β using the asymptotic Gaussian approximation (called the "Wald approximation" in `lme4`) based on the estimated local curvature of the

likelihood surface. Since this approximation is anti-conservative (it assumes infinite data and no model misspecification), we then computed a more conservative parametric bootstrap interval for the intervals that did not include zero. Effects whose interval does not overlap with 0 are significant at the conventional α=0.05 level.

Note that all of the above choices (including the choice of fitting procedure and the power transform of the data) are conservative relative to their alternatives. For instance, prior to power-transforming the duration estimates, the fixed effects of neural pattern distance were estimated to be stronger (as reported in ***Figure 7 – source data 1***.) These alternative analyses revealed additional significant regions that are either false positives or regions we lack the power to detect.

Whole-brain searchlights

In addition to using anatomical ROIs, we ran a cubic searchlight throughout the entire brain. The same analysis as described above was performed for every searchlight, and the Z-value for each searchlight was assigned to the center voxel.

The within-participant analysis was performed in native functional space, and each cubic searchlight contained 3x3x3 (27) voxels. To aggregate the results across participants, each participant's Z-value map was transformed to standard MNI space and down-sampled to 3mm to reflect the resolution of the original data.

The within-interval analysis was performed in 3mm MNI space, in order to match the searchlights across participants. Since this transformation approximately doubles the number

of brain voxels, we ran cubic searchlights of radius 2 with 5x5x5 (125) voxels through the entire brain. Neural pattern distance was not calculated for searchlights on the very edge of the brain with fewer than 25 voxels, in order to reduce noise from overly small patterns. We also excluded a searchlight location if fewer than 5 participants had brain voxels in that location.

Family-wise error rate was controlled using FSL's *randomise* function (version 5.0.4, Winkler, Ridgway, Webster, Smith, & Nichols, 2014). An uncorrected p-value image was first generated, reflecting voxel-wise (searchlight) reliability across participants or intervals. The significance of supra-threshold clusters (defined by the cluster-forming threshold, $p<0.01$) was then assessed by cluster mass. Specifically, a corrected p-value was assigned to each cluster by assessing its cluster mass with respect to the null distribution of the maximum cluster mass during 10,000 permutation simulations (Hayasaka & Nichols, 2003; Nichols & Holmes, 2002). Cluster coordinates are reported in MNI space, and cluster size reflects the number of voxels in 3x3x3mm MNI space.

Comparing speed of pattern change across brain regions

If the brain regions that showed significant effects in our main analysis represent mental context, then the pattern of activity in these regions should change more slowly over time than the patterns in regions representing sensory information. To quantify the speed of pattern change in a given ROI, we obtained the correlation of the pattern at every time point (TR) with itself at every other time point. (As for our main analysis, the BOLD time course of every voxel was smoothed using a moving average filter of 5 TRs. This temporal smoothing was used as a

de-noising technique and did not affect the results.) We then averaged the auto-correlation curves across TRs to obtain a mean auto-correlation function for every region in every participant. The more rapidly a pattern changes over time, the more sharply the auto-correlation should decrease as we move away from 0. To quantify this, we defined the Full-Width Half-Maximum (FWHM) of the auto-correlation curve as the number of time points (TRs) for which the auto-correlation was equal to or greater than half its maximum value (the maximum was always 1.)

To compare the speed of pattern change in the regions we found (right entorhinal cortex and left caudal ACC) with regions involved in auditory and language processing, we performed a paired Wilcoxon signed rank test on the FWHM values across participants. The p-values from this test were subjected to multiple comparisons correction using FDR.

Since the anatomical masks we used varied substantially in size, we sought to ensure that differences in the speed of pattern change were not due to differences in ROI size. For this purpose, we performed the same analysis after regressing the vector of ROI sizes out of the vector of FWHM values for every participant.

Since the above regression would only account for a linear effect of ROI size on the speed of pattern change, we additionally performed a univariate analysis that calculated the auto-correlation function for each voxel individually. The auto-correlation curve was obtained by correlating the BOLD time course of every voxel with itself at all possible lags. The mean auto-correlation for an ROI was obtained by averaging the auto-correlation curves across all the voxels in that ROI. The FWHM values were then calculated in the same manner as above for every ROI in every participant.

Replication of Jenkins and Ranganath 2010 "coarse temporal memory" fMRI analysis

As in Jenkins and Ranganath (2010), we correlated each voxel's activity during encoding of a clip with the accuracy of a participant's placement of that clip on the timeline. Voxel activity was averaged over a 5-TR window centered on the mid-point of the clip. For each participant, the estimated clip position on the timeline was regressed against actual position. Accuracy was defined as the negative error, which was the absolute value of the residual for a clip. Within participants, voxel activity was then correlated with accuracy across all clips, and the Pearson's *r* score was Fisher-transformed. As for the within-participant searchlight analysis, transformed *r* score maps were registered to 3mm MNI space, and FSL's *randomise* was used to control the FWE rate.

Acknowledgments

We would like to thank Lucy Lin for her assistance with data collection for the event boundary experiment. We would like to thank Erez Simony, Lili Sahakyan, Mariam Aly, Anna Schapiro and Michael Chow for their advice on data analysis and preprocessing, as well as helpful discussion.

Prior Presentations, Publications and Co-Authorship

This work was completed in collaboration with Janice Chen, Daniel Toker, Christopher Honey, Michael Shvartsman, Jordan Poppenk, Uri Hasson and Ken Norman. Initial stages of the project have been presented as posters at two conferences as well as a short talk at one meeting:

Lositsky, O., Toker, D., Chen, J., Honey, C.J., Poppenk, J.L., Hasson, U., Norman, K.A. (2013, November). Time perception and contextual drift with a naturalistic stimulus. *Society for Neuroscience Annual Meeting*, San Diego, CA. [***Poster***]

Lositsky, O., Toker, D., Chen, J., Honey, C.J., Poppenk, J.L., Hasson, U., Norman, K.A. (2013, June). Time perception and contextual drift with a naturalistic stimulus. Manhattan Area Memory Meeting, New York, NY. [***Talk***]

Lositsky, O., Toker, D., Chen, J., Honey, C.J., Poppenk, J.L., Hasson, U., Norman, K.A. (2013, May). Time perception and contextual drift with a naturalistic stimulus. Context and Episodic Memory Symposium, Philadelphia, PA. [***Poster***]

The material in this study has also been published at eLife:

Lositsky, O., Chen, J., Toker, D., Honey, C. J., Shvartsman, M., Poppenk, J. L., Hasson, U., & Norman, K. A. (2016). Neural Pattern Change During Encoding of a Narrative Predicts Retrospective Duration Estimates. *eLife*.

References

Andersson, J. L. R., Jenkinson, M., & Smith, S. (2007). *Non-linear registration aka Spatial normalisation FMRIB Technial Report TR07JA2*. *In Practice*. Retrieved from http://fmrib.medsci.ox.ac.uk/analysis/techrep/tr07ja2/tr07ja2.pdf

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*, 1–27. doi:arXiv:1506.04967

Bates, D. M., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221. doi:10.1038/nn1954

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. doi:10.1016/j.neuroimage.2007.04.042

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300. doi:10.2307/2346101

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. doi:10.1093/biomet/93.3.491

Binder, J. R., Frost, J. a, Hammeke, T. a, Bellgowan, P. S., Springer, J. a, Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*(5), 512–528. doi:10.1093/cercor/10.5.512

Block, R. A. (1982). Temporal judgments and contextual change. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *8*(6), 530–544. doi:10.1037/0278-7393.8.6.530

Block, R. A. (1985). Contextual coding in memory: Studies of remembered duration. In *Time, Mind and Behavior* (pp. 169–178).

Block, R. A. (1986). Remembered duration: Imagery processes and contextual encoding. *Acta Psychologica*, *62*(2), 103–122. doi:10.1016/0001-6918(86)90063-6

Block, R. A. (1990). Models of Psychological Time. In *Cognitive models of psychological time* (pp. 1–35).

Block, R. A. (1992). Prospective and retrospective duration judgment: The role of information processing and memory. In F. Macar, V. Pouthas, & W. J. Friedman (Eds.), *Time, actions and cognition: Towards bridging the gap* (pp. 141–152). Dordrecht, The Netherlands: Kluwer Academic.

Block, R. A., & Reed, M. A. (1978). Remembered duration: Evidence for a contextual-change hypothesis. *Journal of Experimental Psychology: Human Learning & Memory*. doi:10.1037/0278-7393.4.6.656

Block, R. A., & Zakay, D. (2008). Timing and remembering the past, the present, and the future. *Psychology of Time*. doi:10.1016/B978-0-08046-977-5.00012-0

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 85–123). Washington, DC: V. H. Winston.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological*, 211–252. doi:10.2307/2287791

Brainard, D. H. (1997, January). The Psychophysics Toolbox. *Spatial Vision*. Brill.

Brown, S. W., & Stubbs, D. A. (1988). The psychophysics of retrospective and prospective timing. *Perception*, *17*(3), 297–310. doi:10.1068/p170297

Bryden, D. W., Johnson, E. E., Tobia, S. C., Kashtelyan, V., & Roesch, M. R. (2011). Attention for Learning Signals in Anterior Cingulate Cortex. *Journal of Neuroscience*. doi:10.1523/JNEUROSCI.4715-11.2011

Buckmaster, C. A., Eichenbaum, H., Amaral, D. G., Suzuki, W. A., & Rapp, P. R. (2004). Entorhinal cortex lesions disrupt the relational organization of memory in monkeys. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *24*(44), 9811–9825. doi:10.1523/JNEUROSCI.1532-04.2004

Carr, V. A., Rissman, J., & Wagner, A. D. (2010). Imaging the Human Medial Temporal Lobe with High-Resolution fMRI. *Neuron*. doi:10.1016/j.neuron.2009.12.022

Chang, C., Cunningham, J. P., & Glover, G. H. (2009). Influence of heart rate on the BOLD signal: The cardiac response function. *NeuroImage*, *44*(3), 857–869.

Chen, J., Honey, C. J., Simony, E., Arcaro, M. J., Norman, K. A., & Hasson, U. (2015). Accessing Real-Life Episodic Information from Minutes versus Hours Earlier Modulates Hippocampal and High-Order Cortical Dynamics. *Cerebral Cortex (New York, N.Y. : 1991)*, 1–14. doi:10.1093/cercor/bhv155

Chen, J., Leong, Y. C., Norman, K. A., & Hasson, U. (2016). Shared experience, shared memory: a common structure for brain activity during naturalistic recall. *bioRxiv*. JOUR. Retrieved from http://biorxiv.org/content/early/2016/01/05/035931.abstract

Chow, M., Chen, J., & Hasson, U. (2015). Latent variable modeling of temporal profiles of neural activity during the processing of continuous natural stimuli. In *Society for Neuroscience Annual Meeting*. Chicago, IL.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models. *Journal of Educational and Behavioral Statistics*, *40*(2), 136–157. doi:10.3102/1076998615570945

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika*, *78*(4), 685–709. doi:10.1007/s11336-013-9328-2

Coull, J. T. (2004). fMRI studies of temporal attention: allocating attention within, or towards, time. *Brain Research. Cognitive Brain Research*, *21*(2), 216–26. doi:10.1016/j.cogbrainres.2004.02.011

Coull, J. T., Vidal, F., Nazarian, B., & Macar, F. (2004). Functional anatomy of the attentional modulation of time estimation. *Science (New York, N.Y.)*, *303*(5663), 1506–8. doi:10.1126/science.1091573

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. doi:10.1016/j.neuroimage.2006.01.021

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. doi:10.1016/j.neuroimage.2010.06.010

Dirnberger, G., Hesselmann, G., Roiser, J. P., Preminger, S., Jahanshahi, M., & Paz, R. (2012). Give it time: neural evidence for distorted time perception and enhanced memory encoding in emotional situations. *NeuroImage*, *63*(1), 591–9. doi:10.1016/j.neuroimage.2012.06.041

Duvernoy, H. M. (2005). *The human hippocampus: functional anatomy, vascularization and serial sections with MRI*. New York: Springer.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123–152. doi:10.1146/annurev.neuro.30.051606.094328

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological*

Science : A Journal of the American Psychological Society / APS, *22*, 243–252. doi:10.1177/0956797610393742

Ezzyat, Y., & Davachi, L. (2014). Similarity breeds proximity: Pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron*, *81*, 1179–1189. doi:10.1016/j.neuron.2014.01.042

Faber, M., & Gennari, S. P. (2015). In search of lost time: Reconstructing the unfolding of events from memory. *Cognition*, *143*, 193–202. doi:10.1016/j.cognition.2015.06.014

Ferstl, E. C., Neumann, J., Bogler, C., & Von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, *29*(5), 581–593. doi:10.1002/hbm.20422

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11050–11055. doi:10.1073/pnas.200033797

Fischl, B., Liu, A., & Dale, A. M. (2001). Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, *20*(1), 70–80. doi:10.1109/42.906426

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., … Dale, A. M. (2004). Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, *14*(1), 11–22. doi:10.1093/cercor/bhg087

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, *10*(4), 507. doi:10.2307/2331838

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790942

Gurka, M. J., Edwards, L. J., Muller, K. E., & Kupper, L. L. (2006). Extending the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *169*(2), 273–288. doi:10.1111/j.1467-985X.2005.00391.x

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313. doi:10.1016/j.tics.2015.04.006

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science (New York, N.Y.)*, *303*(5664), 1634–1640. doi:10.1126/science.1089506

Hasson, U., Nusbaum, H. C., & Small, S. L. (2007). Brain networks subserving the extraction of sentence information and its encoding to memory. *Cerebral Cortex*, *17*(12), 2899–2913.

doi:10.1093/cercor/bhm016

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(10), 2539–2550. doi:10.1523/JNEUROSCI.5487-07.2008

Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: Random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356. doi:10.1016/j.neuroimage.2003.08.003

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1–2), 67–99. doi:10.1016/j.cognition.2003.10.011

Hicks, R. E., Miller, G. W., & Kinsbourne, M. (1976). Prospective and retrospective judgments of time as a function of amount of information processed. *The American Journal of Psychology*, *89*(4), 719–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1020767

Hindy, N. C., & Turk-Browne, N. B. (2015). Action-Based Learning of Multistate Objects in the Medial Temporal Lobe. *Cerebral Cortex*, 1–13. doi:10.1093/cercor/bhv030

Howard, M. W., & Eichenbaum, H. (2013). The hippocampus, time, and memory across scales. *Journal of Experimental Psychology. General*, *142*(4), 1211–30. doi:10.1037/a0033621

Howard, M. W., Fotedar, M. S., Datey, A. V, & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychological Review*, *112*(1), 75–116. doi:10.1037/0033-295X.112.1.75

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. doi:10.1006/jmps.2001.1388

Jacobs, N. S., Allen, T. A., Nguyen, N., & Fortin, N. J. (2013). Critical role of the hippocampus in memory for elapsed time. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(34), 13888–93. doi:10.1523/JNEUROSCI.1733-13.2013

Jenkins, L. J., & Ranganath, C. (2010). Prefrontal and medial temporal lobe activity at encoding predicts temporal context memory. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*, 15558–15565. doi:10.1523/JNEUROSCI.1337-10.2010

Jenkins, L. J., & Ranganath, C. (2016). Distinct neural mechanisms for remembering when an event occurred. *Hippocampus*, n/a-n/a. doi:10.1002/hipo.22571

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. M. (2002). Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*, 825–841. doi:10.1006/nimg.2002.1132

Kurby, C. a, & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–9. doi:10.1016/j.tics.2007.11.004

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*, 2906–2915. doi:10.1523/JNEUROSCI.3684-10.2011

Lipton, P. A., & Eichenbaum, H. (2008). Complementary roles of hippocampus and medial entorhinal cortex in episodic memory. *Neural Plast*, *2008*, 258467. doi:10.1155/2008/258467

Lipton, P. A., White, J. A., & Eichenbaum, H. (2007). Disambiguation of overlapping experiences by neurons in the medial entorhinal cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(21), 5787–5795. doi:10.1523/JNEUROSCI.1063-07.2007

Livesey, A. C., Wall, M. B., & Smith, A. T. (2007). Time perception: manipulation of task difficulty dissociates clock functions from other cognitive demands. *Neuropsychologia*, *45*(2), 321–31. doi:10.1016/j.neuropsychologia.2006.06.033

Maguire, E. A., Frith, C. D., & Morris, R. G. M. (1999). The functional neuroanatomy of comprehension and memory: The importance of prior knowledge. *Brain*, *122*(10), 1839–1850. doi:10.1093/brain/122.10.1839

Manning, J. R., Kahana, M. J., & Norman, K. A. (2014). The role of context in episodic memory. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The Cognitive Neurosciences* (V., pp. 557–566). Cambridge, MA: MIT Press.

Manns, J. R., Howard, M. W., & Eichenbaum, H. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, *56*(3), 530–40. doi:10.1016/j.neuron.2007.08.017

Mar, R. A. (2004). The neuropsychology of narrative: Story comprehension, story production and their interrelation. *Neuropsychologia*. doi:10.1016/j.neuropsychologia.2003.12.016

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron*, *84*(4), 870–881. doi:10.1016/j.neuron.2014.10.013

Mensink, G.-J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*. doi:10.1037/0033-295X.95.4.434

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. doi:10.1002/hbm.1058

Noulhiane, M., Pouthas, V., Hasboun, D., Baulac, M., & Samson, S. (2007). Role of the medial temporal lobe in time estimation in the range of minutes. *Neuroreport*, *18*(10), 1035–8. doi:10.1097/WNR.0b013e3281668be1

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 1–10. doi:10.1073/pnas.1305373110

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–42.

Pollatos, O., Laubrock, J., & Wittmann, M. (2014). Interoceptive focus shapes the experience of time. *PLoS ONE*, *9*(1). doi:10.1371/journal.pone.0086934

Polyn, S. M., & Kahana, M. J. (2008). Memory search and the neural representation of context. *Trends in Cognitive Sciences*, *12*(1), 24–30. doi:10.1016/j.tics.2007.10.010

Poppenk, J., & Norman, K. a. (2014). Briefly cuing memories leads to suppression of their neural representations. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(23), 8010–20. doi:10.1523/JNEUROSCI.4584-13.2014

Poppenk, J., & Norman, K. a. (2012). Mechanisms supporting superior source memory for familiar items: A multi-voxel pattern analysis study. *Neuropsychologia*, *50*(13), 3015–3026. doi:10.1016/j.neuropsychologia.2012.07.010

Poynter, W. D. (1983). Duration judgment and the segmentation of experience. *Memory & Cognition*, *11*, 77– 82.

Ritchey, M., & Ranganath, C. (2012). Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience*. doi:10.1038/nrn3338

Sahakyan, L., & Smith, J. R. (2014). "A long time ago, in a context far, far away": retrospective time estimates and internal context change. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(1), 86–93. doi:10.1037/a0034250

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627. doi:10.1016/j.cub.2012.06.056

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, *22*(3), 1060–1075. doi:10.1016/j.neuroimage.2004.03.032

Segonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically Accurate Topology-Correction of Cortical Surfaces Using Nonseparating Loops. *IEEE Transactions on Medical Imaging*, *26*(4), 518–529. doi:10.1109/TMI.2006.887364

Shapleske, J., Rossell, S. ., Woodruff, P. W. ., & David, A. . (1999). The planum temporale: a systematic, quantitative review of its structural, functional and clinical significance. *Brain Research Reviews*, *29*(1), 26–49. doi:10.1016/S0165-0173(98)00047-2

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–40. doi:10.1016/j.neuron.2013.07.007

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(43), E4687–E4696. doi:10.1073/pnas.1323812111

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, *17*(1), 87–97. doi:10.1109/42.668698

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. doi:10.1002/hbm.10062

Stephens, G. J., Honey, C. J., & Hasson, U. (2013). A place for time: the spatiotemporal structure of neural dynamics during natural audition. *Journal of Neurophysiology*, *110*(9), 2019–26. doi:10.1152/jn.00268.2013

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J. D., & Doyne Farmer, J. (1992). Testing for Nonlinearity in Time-Series -- the Method of Surrogate Data. *Physica D*, *58*(1–4), 77–94. doi:10.1016/0167-2789(92)90102-S

Wiener, M., Turkeltaub, P., & Coslett, H. B. (2010). The image of time: A voxel-wise meta-analysis. *NeuroImage*, *49*(2), 1728–1740. doi:10.1016/j.neuroimage.2009.09.064

Wilson, D. I. G., Langston, R. F., Schlesiger, M. I., Wagner, M., Watanabe, S., & Ainge, J. A. (2013). Lateral entorhinal cortex is critical for novel object-context recognition. *Hippocampus*, *23*(5), 352–366. doi:10.1002/hipo.22095

Wilson, D. I. G., Watanabe, S., Milner, H., & Ainge, J. A. (2013). Lateral entorhinal cortex is necessary for associative but not nonassociative recognition memory. *Hippocampus*, *23*(12), 1280–1290. doi:10.1002/hipo.22165

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397. doi:10.1016/j.neuroimage.2014.01.060

Wittmann, M. (2013). The inner sense of time: how the brain creates a representation of duration. *Nature Reviews. Neuroscience*, *14*, 217–23. doi:10.1038/nrn3452

Wittmann, M., Simmons, A. N., Aron, J. L., & Paulus, M. P. (2010). Accumulation of neural

activity in the posterior insula encodes the passage of time. *Neuropsychologia*, *48*, 3110–3120. doi:10.1016/j.neuropsychologia.2010.06.023

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in Reading and Film Comprehension, *138*(2), 307–327. doi:10.1037/a0015305

Zakay, D., & Block, R. A. (2004). Prospective and retrospective duration judgments: An executive-control perspective. In *Acta Neurobiologiae Experimentalis* (Vol. 64, pp. 319–328).

Zakay, D., Tsal, Y., Moses, M., & Shahar, I. (1994). The role of segmentation in prospective and retrospective time estimation processes. *Memory & Cognition*, *22*(3), 344–351. doi:10.3758/BF03200861

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. doi:10.1037/1082-989X.12.4.399

Study 2: Weak Fusion of Cue Predictions in Context-Based Decision-Making

Introduction

Background

Surprise is relative. We may be very surprised to see our colleague walking around the grocery store over the weekend, but perfectly prepared to see them in the laboratory, where we had just been formulating a question to ask them. In this manner, contextual information (such as our current location or goal) can predict which events are more likely or which actions might be relevant. However, in everyday life, we often face a trade-off between learning general expectations ("priors") that apply to several contexts (e.g. it is *always* important to be prepared for scientific discussions) and learning priors for specific contexts that help us adapt to particular situations (e.g. scientific discussions are not frequent in the grocery store, whereas food shopping is). On the one hand, learning general expectations can help us make predictions about new, previously unseen situations. Moreover, general priors might require fewer resources (require storing and updating only a single probability rather than a separate probability for each situation) and help to average noisy observations across many different contexts. On the other hand, if the contexts truly do make very different predictions, learning the probabilities for a specific context could gain a lot of time and value.

How do humans resolve this trade-off? Previous work has shown that people prefer simpler explanations of their observations, unless the provided evidence warrants complex explanations (Genewein & Braun, 2014; Pacer & Lombrozo, 2017). In our study, a "simple" explanation of the data would amount to assuming that a single distribution governs stimulus probabilities across all contexts, and therefore averaging experienced probabilities across contexts. Additionally, multiple studies of perceptual decision-making, reviewed below, have suggested people combine predictions across perceptual cues, unless there is sufficient evidence that the cues make different predictions (Gershman & Niv, 2013; Girshick & Banks, 2009; Knill, 2007).

In the domain of perception, a rich literature exists on how people combine information both within and across sensory modalities to determine a property of an object. For instance, how are sensory cues (such as binocular disparity, object motion, occlusion) combined to determine a single property, such as the depth of an object in a visual scene? Multiple studies have suggested that when the various cues provide slightly different information, organisms average depth information across cues, but weight each cue by its reliability (which could vary with trial by trial noise, or could be determined by previous history of how informative a cue is in a certain context; Ernst & Banks, 2002; Gepshtein & Banks, 2003). Moreover, some studies have found that humans assign increasingly less weight to a cue as it begins to conflict with the other cues (Girshick & Banks, 2009; Knill, 2007). Such results are in line with theories that cue integration should exhibit "robust estimation": as a cue provides more and more different information from other cues, its effect on the percept should decrease, because its likelihood of being generated by the same "cause" (the same source) is decreasing. A system with such

properties should exhibit non-monotonic changes of the percept as a function of changes in the conflicting cue: the percept should vary with the cue for small discrepancies, but reverse direction for large discrepancies (Landy, Maloney, Johnston, & Young, 1995).

Although the experimental paradigms used in cue combination studies are very different from the decision making task we will use here, and although the neural mechanisms underlying sensory cue integration are likely very different from those underlying context-dependent decisions, the computational principle is the same. In fact, Hospedales & Vijayakumar (2009) frame the challenges of cue integration as a causal inference problem: how likely are the two cues to be generated by the same cause (the same percept)? It is hypothesized that organisms integrate cues to the degree that they are generated by the same percept, but discard cues to the degree that they appear to be outliers or to stem from a different source.

Gershman and Niv (2013) tested whether these principles of Bayesian inference can explain whether predictions from different perceptual attributes (e.g., two colors of dots) are integrated to generate people's priors about the number of circles on a screen. In their study, participants estimated the number of circles on the screen (presented briefly enough to prevent counting), and all the dots on a given trial appeared in one of two colors. In half of the blocks, the two colors predicted highly overlapping distributions of circle numbers (e.g. mean of 65 vs. 55 circles), while in the other half of blocks, the colors predicted more distinct distributions (e.g., mean of 65 vs. 35 circles). They hypothesized that blocks with overlapping distributions would encourage "merging" of the circle number probabilities across colors, while blocks with more distinct ones would encourage "categorization" of circle number probability

by color, such that people might form distinct expectations for each color. Across three experiments, they found confirmation of their hypothesis: people's estimates for the two colors were *more* biased towards the average of the two distributions when they were more similar, and *less* biased towards the average when they were more distinct. Moreover, people's estimates for the two colors were correlated across trials when their distributions were similar, suggesting that their predictions were being combined.

Based on these two lines of work, we hypothesized that humans might group expectations across contexts when the contexts make sufficiently similar predictions about a particular event. For instance, if the probability of seeing my colleague in the lab were only a little higher than at the grocery store, I might average the probability of seeing my colleague across those two environments. However, if the grocery store predicts a far lower probability of seeing my colleague, then I might form separate expectations for these two contexts.

In the above example, the context (grocery store vs. lab) predicted both which stimulus might occur (e.g., my colleague) and which action I might respond with (e.g., scientific discussion). However, context may not only influence which stimuli we might see and which responses we might make independently, but also their interaction — that is, which responses are appropriate to which stimuli. For instance, discussing a research project may be an ideal response to seeing my colleague in the lab, but may not be as appropriate when running into them in the grocery store. Instead, discussing our upcoming departmental retreat could be more appropriate in the grocery store. Thus, even if both the lab and grocery store predict the same probability of seeing my colleague, they might predict different probabilities of performing certain actions (such as discussing our research project). In such cases, can people

group the probability of seeing their colleague across the two contexts, yet still form separate predictions about the appropriate responses?

While the above studies investigated how predictions are grouped across cues as a function of their similarity (Gershman and Niv, 2013; Girshick & Banks, 2009), in those experiments, predictions about stimuli could not be dissociated from predictions about the appropriate responses, as the stimuli required precisely the same responses regardless of context. In such experiments, once the cues had been merged to form a unified prediction, individual cue information could often be discarded without any loss in performance. In fact, in the literature on perceptual cue combination, there is debate about whether the system retains information about each individual feature, such that behavior could be governed both by the unified percept on some trials, and by individual cue information on other trials. In Landy et al. (1995), integration mechanisms that retain individual feature information are termed "weak fusion", while those that only retain the resulting unified percept are termed "strong" or "mandatory fusion". One study (Hillis, Ernst, Banks, & Landy, 2002) found evidence that information about individual cues is lost when they are integrated *within* a sensory modality, such as vision, but retained and used together with the unified percept when the cues span multiple modalities (vision and touch). Hospedales & Vijayakumar (2009) interpret this as further evidence that cue combination obeys principles of latent cause inference: weak fusion across senses would be adaptive if cues from different modalities are less likely to be generated by the same source.

Still, in most of the perceptual decision-making paradigms reviewed above, the task could be performed correctly even if information about individual cues was lost after integration, and performance could even improve by using only the unified percept to guide behavior (Ernst &

Bülthoff, 2004). Thus, it is unclear whether predictions will still be merged across contextual cues in context-dependent decisions, in which differentiating contexts is crucial for responding appropriately. Our study set out to explore whether such merging across contexts would occur and whether it would impair people's ability to act in a context-appropriate manner.

Our approach

We investigated this question using a task called the AX-CPT, in which the correct response to a stimulus depends on a cue that appears seconds before it. In this task, the cue (A or B) serves as the "context" for the stimulus (X or Y), both because it allows the participant to prepare for the upcoming stimulus and because it determines which action is correct for each probe. Specifically, participants were instructed to press the left button after AX and BY pairs, but to press the right button after AY and BX pairs (**Figure 14**). Since the cue was always necessary for making the correct response, we were able to test whether participants could merge predictions about the response across cues, while still differentiating the appropriate stimulus-response rules by cue. To do this, we parametrically varied the probability of the stimulus given the cue across participants: for some participants (MERGE group), both cues predicted similar response probabilities (left was correct more frequently than right) but different stimulus probabilities; for other participants (SPLIT group), the cues predicted different response probabilities, but similar stimulus probabilities (X was more frequent than Y). We expected that participants would use the cue to prepare the most likely response, especially in the SPLIT group, when the two cues predicted sufficiently different responses, but

that they would combine predictions about the response across cues when they were similar enough, developing a general response prior that disregards context.

To test this hypothesis, we built a variant of the drift diffusion model (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff, 1978) in which the response rule (press left for X vs. press right for X) depends on the cue. In addition to influencing the appropriate response, the cue could also influence the starting point for the decision, reflecting the participant's prior about which response is more likely given the cue they just saw. Fitting the model to accuracy and reaction times on the AX-CPT task enables us to infer the degree to which participants prepared the left or right response given the cue, and to test how this preparation differs as a function of experimental condition. In the first part of the Results, we report how the relationship between A and B response priors varies as a function of experimental condition ("Implicit Response Priors" section).

Importantly, in addition to analyzing participants' "implicit" priors based on their reaction times and accuracy, we also asked participants to explicitly estimate the proportion of each trial type (A-X, A-Y, B-X and B-Y) during debriefing ("Explicit Response Priors" section). This enabled us to test whether any sub-optimality in response priors measured during task performance could be explained by inaccurate perception of cue-stimulus probabilities. Specifically, we tested whether merging of response probabilities across cues would be visible in explicit probability estimates, which would suggest that merging affects the learning and perception of probabilities, and does not occur solely during the use of these probabilities to prepare decisions.

Finally, correlating participants' explicit A estimates with their explicit B estimates revealed an additional pattern we had not anticipated. As mentioned above, when A and B predicted different response probabilities (left for A, right for B) they nevertheless predicted similar stimulus probabilities (X more frequent than Y). In this condition (SPLIT group), we predicted that participants would learn to use the cue to prepare their response and develop separate response priors for each cue. While the latter was confirmed, we also observed that their estimates now correlated along the X/Y dimension. In other words, while participants did develop different response probabilities for the two cues, they now appeared to average ("merge") stimulus probabilities across cues.

To verify that this averaging reflected their internal representation of the task (rather than reflecting correlations in the environment), we ran a follow-up study in which the probabilities of X and Y for B cues changed in the middle of the experiment. This study showed that participants continued to group their observations according to the same dimension they learned earlier in the experiment, even when the environment no longer favored such groupings.

Our results reveal a remarkable flexibility in the way human participants group their observations across contexts. Participants seem able to merge response predictions across cues to the degree that they are similar, while still differentiating the correct stimulus-response rules by cue. Moreover, they seem able to merge stimulus probabilities across cues when these are similar, yet still able to form separate predictions about the appropriate response.

Methods Summary

Summary of Experimental Manipulation

In order to study how people use cues to prepare a behavioral response, we used the AX-CPT, a task in which the correct response to a stimulus depends on a cue that appeared seconds earlier. On every trial, a cue (A or B) appeared two seconds before a probe stimulus (X or Y). Participants are instructed to press the left button after AX and BY pairs, but to press the right button after AY and BX pairs [**Figure 14**].

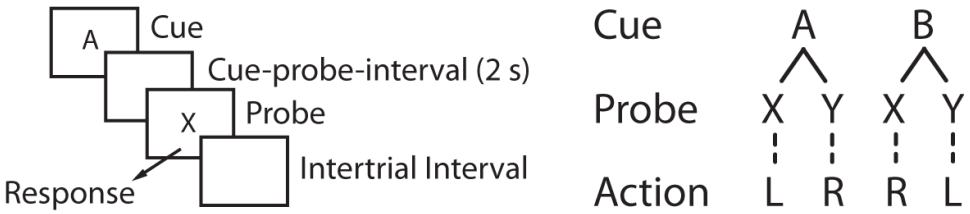


Figure 14 The AX-CPT task. On every trial, a cue (A or B) determines the correct response rule for the probe (X or Y), which appears 2 seconds later. Participants are asked to press the left or right button as soon as the probe appears and are given 5 seconds to respond. They are rewarded a small sum of money for each correct response (the exact amount differed between lab and M-turk participants).

To test in which situations people form specific expectations for each cue, participants were given this task with different probabilities of seeing X and Y depending on the cue. The two cues were equally frequent, but the probability of the probe given the cue varied across participants. For A cues, all participants saw X more frequently than Y (and therefore pressed the left button more). For B cues, we varied the probability of seeing X (as opposed to Y) parametrically across

participants. Some participants saw more X's (and therefore pressed the right button more) while other participants saw more Y's (and pressed the left button more). [**Figure 15**]

We hypothesized that when A and B cues both predict the left response more than the right response (when AX and BY trials are more frequent than AY and BX trials), participants would form a general expectation to press left, grouping the response probabilities across cues, even when they would perform better by learning the specific response probabilities for each cue. For instance, when BX=20% and BY=30% (third condition in **Figure 15**), A predicts the left response more strongly than B, such that better performance (both speed and accuracy) can be achieved by distinguishing the degree of left bias between A and B. However, a participant with a preference for simpler representations (such as Agent 2 above) might nevertheless average them, sacrificing reward rate in the service of simplicity. At the same time, when A and B make sufficiently different predictions about the response, such that A predicts more frequent left responses (all participants) while B predicts more frequent right responses (when AX and BX trials are more frequent), we expected participants to learn separate response priors for A and B cues.

MERGE group

AX	40%	Left
AY	10%	Right
BX	10%	Right
BY	40%	Left

MERGE group

AX	40%	Left
AY	10%	Right
BX	15%	Right
BY	35%	Left

MERGE group

AX	40%	Left
AY	10%	Right
BX	20%	Right
BY	30%	Left

AX	40%	Left
AY	10%	Right
BX	25%	Right
BY	25%	Left

SPLIT group

AX	40%	Left
AY	10%	Right
BX	30%	Right
BY	20%	Left

SPLIT group

AX	40%	Left
AY	10%	Right
BX	35%	Right
BY	15%	Left

SPLIT group

AX	40%	Left
AY	10%	Right
BX	40%	Right
BY	10%	Left

Figure 15 Experimental design. Each participant received one set of probabilities that did not change throughout the experiment. A and B cues appeared equally often, but the probability of the probe (X / Y) given the cue varied across participants. The frequencies of AX and AY trials were always 40% and 10%, such that all participants pressed the left button more often after A. The frequencies of BX and BY varied parametrically across participants. For MERGE groups, BY was more common than BX, such that they pressed the left button more after B. We expected that these participants would merge the probability of the left response across cues, since the two cues predicted similar response probabilities. For SPLIT groups, BX was more common than BY, such that they pressed the right button more after B. We expected that these participants would learn that A and B predict different response probabilities and form separate priors for each cue. All participants were told the rules of the task (press left after AX and BY, press right after AY and BX), but not informed that some trials were more frequent than others. At debriefing, they were asked to estimate the frequencies of each of the four trial types.

Participant Populations

To increase statistical power, and evaluate whether our results generalize across populations, we collected behavioral data both in the lab and on Amazon's Mechanical Turk (M-turk). As mentioned in the Methods, we collected three groups of participants, which differed in the following ways.

The "Lab 1" group received 480 trials, after which they were asked to provide frequency estimates during debriefing. For Lab 1 participants, only the letters "A" and "B" were used as

cues and only the letters “X” and “Y” were used as probes, but the frequencies assigned to each letter were roughly counter-balanced across participants. The “D” and “L” keys on the keyboard constituted the left and right response buttons (the frequencies assigned to each button were counter-balanced). Although trial order was pseudo-randomized, the trial frequencies assigned to a participant were constrained to be precise within blocks of 20 trials. For instance, if a participant received the 40-10-10-40 frequencies (MERGE group, Figure 2), the four trial types appeared with exactly those frequencies within a block of 20 trials.

The “Lab 2” and “M-turk” groups both received 240 trials, after which they were asked to provide frequency estimates during debriefing. For both of these groups, any consonant in the alphabet could be randomly chosen as a cue or probe (e.g. “L”, “T”, “J” and “C” could have been the chosen letters for a participant.) The “4” and “8” keys constituted the left and right response buttons (again, the frequencies assigned to each button were counter-balanced). For these two groups, the trial frequencies assigned to a participant were constrained to be precise within blocks of 40 trials.

Finally, task rules (press left for AX and BY, press right for AY and BX) were presented in slightly different ways between the M-turk group and the two lab groups. M-turk participants saw the A and B rules on separate screens, and practiced the rule for each cue for 6 trials before seeing the rule for the second cue (the order of cue presentation was counter-balanced). They then had to practice the rules for all four trial types and respond accurately 10 trials in a row before proceeding to the actual task. In contrast, since we were less concerned about the data quality of in-lab participants, lab participants saw the correct responses for all four trial types on the same screen (the order of presentation was counter-balanced), and

proceeded immediately to the task. While we did not believe this difference could significantly impact performance, it is possible that some of the difference between M-turk and lab participants could have arisen from this difference in the instruction phase.

Implicit and Explicit Measures of Response Priors

In order to quantify participants' expectation of the response depending on the cue, we used an implicit measure (participants' reaction times and accuracy on left vs. right responses after A and B) as well as an explicit measure, asking them to estimate the frequency of each trial type (AX, AY, BX and BY) at the end of the experiment.

The simple intuition behind the implicit measure is that, if participants expect to press left more frequently than right after A cues, they should be faster and more accurate on left than right responses after A. Beyond analyzing raw reaction time and accuracy data, we built a new variant of the drift diffusion model (Bogacz et al., 2006; Ratcliff, 1978) in which the decision to press left or right depends on context. In the model, the distance between the starting point and threshold for the decision quantifies people's expectation of pressing the left or right button before the probe appears (Leite & Ratcliff, 2011). We allowed these parameters to vary depending on the cue and defined the A prior as the ratio between the best-fit starting and best-fit threshold on A trials (likewise for the B prior). We then analyzed how the difference between the best-fit A and B priors varied as a function of the true difference in response probabilities between A and B.

Fitting the model confers several advantages over analyzing raw behavioral data. First, fitting the threshold parameter allows us to account for any speed-accuracy trade-offs in

behavior. Second, since the fitting procedure uses both accuracy and the entire distribution of reaction times to estimate the best-fit parameters, it allows us to infer participants' priors from both measures simultaneously.

Model Predictions under Different Strategies

In addition to providing a more precise method for analyzing behavior, the model allows us to make predictions about how an "optimal" (or bounded optimal) agent would behave under different constraints. Throughout the results section, we test which of three qualitatively different strategies provides the best explanation of behavior on both the implicit (reaction times and accuracy) and explicit (frequency estimates) measures: 1) an agent that always learns separate priors for A and B cues and never combines predictions across cues, 2) an agent that always groups response predictions across cues, regardless of their similarity, and 3) an agent that groups response predictions across cues only when they are similar enough, but learns separate response probabilities for A and B when they are sufficiently different. While Agent 3 was the most consistent with our hypothesis based on previous findings, we wanted to rule out the possibility that either of the two extremes (*never* grouping priors across contexts (Agent 1) or *always* grouping priors across contexts (Agent 2)) was a more parsimonious explanation of human behavior.

To predict how behavior might look under these three strategies, we computed the optimal starting points and thresholds for both A and B cues using three different "optimization functions" corresponding to the three types of Agents described above: 1) Agent 1 seeks only to maximize reward rate, 2) Agent 2 seeks to maximize reward but also minimize the difference

between A and B priors, and 3) Agent 3 seeks to maximize reward rate but also minimize the difference between A and B priors *to the degree that A and B make similar predictions*.

More formally, the three agents find the starting points and thresholds that maximize the expressions below:

Agent 1 $RewardRate$

Agent 2 $RewardRate - \frac{\lambda}{1-\lambda} * |\, APrior - BPrior \,|$

Agent 3 $RewardRate - \frac{\lambda}{1-\lambda} * |\, APrior - BPrior \,| * ABsimilarity$

$ABsimilarity = |\, AX\% - BX\% \,|$

where λ is the cost of separating A and B priors, meant to symbolize a bias towards simpler representations, and ABsimilarity is none other than the experimental condition (as BX % increases, the B cue begins to predict the right response more and its predictions diverge from those of the A cue).

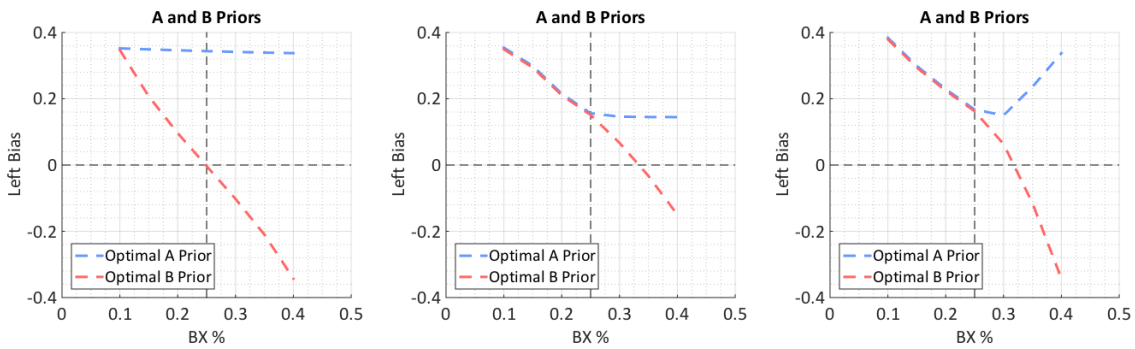


Figure 16 Optimal A and B priors under three different strategies. The optimal priors are the starting point / threshold ratios for A and B cues given fixed settings of the other parameters (drift rate, cue noise, T0) and a certain optimization function. **Agent 1** (left panel) seeks only to maximize reward rate; it treats A and B priors as separate, so long as the cues differ even slightly in their predictions. **Agent 2** (middle panel) seeks to maximize reward and minimize the difference between A and B priors; it groups priors across cues regardless of the similarity of their predictions, such that A and B priors are always closer to each other than "optimal" (Agent 1). **Agent 3** (right panel) seeks to maximize reward rate and minimize the difference between A and B priors to the degree that A and B make similar predictions; it groups priors across cues to the degree that they make similar predictions, but separates priors by cue when they are different enough. The y-axis (degree of left bias) reflects the degree to which the starting point is close to the left decision boundary; negative values indicate starting points closer to the right boundary. The x-axis (% BX trials) reflects the experimental condition that a participant might receive. Since all participants saw 40% AX trials and 10% AY trials, the ratio of left:right responses for A cues was always 4:1; this is why the optimal A prior is constant across conditions for Agent 1 (which treats A independently from B). On the other hand, the ratio of BX:BY trials varied between 1:4 and 4:1 across participants; this is why the optimal B prior decreases from left-biased to right-biased across conditions for Agent 1.

Figure 16 shows the optimal A and B priors as a function of experimental condition for the three agents. When the proportion of BX trials is 10% (left-most condition in each plot), both A and B predict that the left response is 4 times more likely than the right response. For this condition, all three agents should adopt very similar priors for A and B cues. However, as the proportion of BX trials increases to 25%, B begins to predict the left response *less* than A: in those conditions, the "optimal" Agent 1, which treats the two cues independently, begins to adopt separate response priors for A and B, since adapting to the specific probability of left

after each cue earns more reward rate. On the other hand, agents 2 and 3, which are pressured to group response priors across cues, continue to adopt the same value for both A and B priors, since their predictions about the left response are not different enough: having separate priors would not earn enough additional reward rate to offset the cost of splitting the representations. Finally, when the proportion of BX trials increases above 25%, the B cue predicts the right response more than the left. In those conditions, all three agents begin to separate response priors by cue, because they can earn more reward rate by having a leftward prior on A and a rightward prior on B. However, note that Agent 3, because it groups priors across cues only when they are similar enough, at some point begins to treat A and B priors as independent; in other words, it begins to resemble Agent 1 when A and B predictions are different enough. At the same time, Agent 2 continues to average response predictions across the two cues even when they are most different, such that its A and B priors are closer to each other than those predicted by Agents 1 or 3.

To summarize, Agent 3 behaves as a hybrid between Agents 1 and 2, grouping response priors across cues when they make similar predictions, and separating them when they make different predictions. This two-mode strategy results in a U-shaped curve of the A prior as a function of B.

Unfortunately, we were prevented from fitting these models directly to human behavior and comparing their goodness of fit for two reasons. First, we found an additional pattern in the behavior that was partially confounded with merging priors across cues: participants' best-fit priors were always closer to 0 than those predicted by any of the three agents (this is described in detail below). Second, the three models can only be differentiated when

comparing behavioral across all frequency conditions, since some of the conditions could look the same under all models (for instance, all three make identical predictions when BX%=10%, Agents 1 and 3 behave similarly when BX%=40%, and Agents 1 and 2 behave similarly when BX%<25%). Since each participant only received one of the 7 conditions (7 sets of probabilities shown in Figure 2), separate parameters and cost functions had to be estimated for each participant and the likelihood surface for the whole population became unstable (see Methods for more details).

Since we could not precisely compare goodness of fit, we leveraged the fact that the three models make qualitatively different predictions about how the prior for A should vary as a function of the prior for B. We therefore tested whether B statistics had an influence on the A prior (Agents 2 and 3 vs. Agent 1) and if so, whether the A prior varied as a function of B in a monotonic (Agent 2) or non-monotonic (Agent 3) manner. We examined this using both the implicit and explicit behavioral measures of frequency estimation.

Results

Outline of Results

We first present an outline of the results, exposing how the main conclusions are organized into sections. The results are organized by two motivating questions: investigating how participants combine probabilities across cues as a function of the similarity between cue predictions, and investigating whether the priors participants use for their decisions are optimal with respect to their probability estimates. All the sections leading to "Experiment 2" aim to address the first question, while all the sections after "Experiment 2" aim to address the second question.

Our analyses focused on two measures of participants' expectations given the cue: 1) participants' *implicit* response priors, as measured by model fits to their reaction time and accuracy data, and 2) participants' *explicit* response priors, as measured by their estimates of the trial frequencies during debriefing. (See the "Implicit and Explicit Measures of Response Priors" section.)

For both of these measures, we tested which of three qualitative models was a better description of how participants combined priors across cues: 1) an agent that always learns separate priors for A and B cues and never combines predictions across cues, 2) an agent that always groups response predictions across cues, regardless of their similarity, and 3) an agent that groups response predictions across cues only when they are similar enough, but learns

separate response probabilities for A and B when they are sufficiently different. (See the "Model Predictions under Different Strategies" section.)

To characterize the relationship between A and B priors, we first used linear regression models to test how participants' priors on A cues varied as the frequencies of BX/BY trials changed across participants, both for the implicit and explicit measures ("Implicit Response Priors" and "Explicit Response Priors" sections). Next, we examined whether individual differences in the B prior across participants predicted individual differences in the A prior, for both implicit and explicit measures. In other words, if a participant's B prior was right-biased, did their A prior tend to be right-biased as well? (See the "Correlation between A and B Priors Varies by Similarity Between Cue Predictions" section.)

Finally, we tested whether participants who experienced similar vs. different response probabilities for A and B had learned different ways of grouping their observations across cues by performing a follow-up experiment, in which the trial frequencies changed half-way through the task ("Experiment 2: Changing Trial Frequencies Within Participants" section).

Having explored how probabilities were combined across cues as a function of their similarity, all subsequent analyses investigated the relationship between participants' implicit priors (measured by their reaction times and accuracy on the task) and their explicit estimates of the probabilities at debriefing ("Relating Explicit and Implicit Measures: Do distortions in probability estimates account for sub-optimality in decision-making?" section). First, we quantified the correlation between the two measures across participants using a regression model, testing whether a participant's explicit estimates of AX:AY probabilities could predict

their implicit A prior ("Participants' Explicit Estimates Predict Their Implicit Priors" section.) Second, we tested whether using a participant's estimated frequencies in the model, rather than the true stimulus frequencies, significantly improved model predictions about their behavior on the task ("Estimated frequencies improve model fits relative to true frequencies" section). Third, we characterized the ways in which participants' implicit priors differed from the optimal ones predicted by the model given their frequency estimates ("Best-fit priors deviate from optimal ones in systematic ways" section). This section revealed that averaging probabilities across cues was not the only way in which participants deviated from optimality. Another consistent pattern in the behavior was that participants used overly small starting points, such that they were less left-biased than optimal when left was more frequent, and less right-biased than optimal when right was more frequent ("Implicit and Explicit Priors are Closer to 0 than Optimal" section). We conclude by stating that a model aiming to describe participants' priors on this task needs to account for both of the patterns we found: merging probabilities across cues and underestimating probability ratios ("Requirements for a model that could capture participants' implicit priors" section.)

The last section of the Results – "Alternative Interpretations and Additional Analyses" – contains additional analyses that help shed light on the mechanisms behind merging probabilities across cues. This includes analyses of how overall task performance changes as a function of merging ("Cue noise, error rates and reaction times as a function of condition"), whether merging appears to be an "all or nothing" phenomenon ("Merging appears graded, rather than binary, both across and within participants"), analyses aiming to test whether

stimulus probabilities are being merged across cues ("Averaging stimulus instead of response probabilities") and how switch costs change as a function of trial frequencies ("What role do switch costs play in our implicit measure?").

Experiment 1: Changing Trial Frequencies Across Participants

Implicit Response Priors

Drift Diffusion Model Fits to Experimental Data

To evaluate which of the three strategies above best describes observed behavior, we first assessed participants' *implicit* response priors by fitting the drift diffusion model to their reaction time and accuracy data. The best-fit starting point / threshold ratio on A trials was defined as the A prior and the best-fit starting / threshold ratio on B trials was defined as the B prior. **Figure 17** shows the best-fit A and B priors as a function of experimental condition for the three participant groups, plotted using the same conventions as in **Figure 16**.

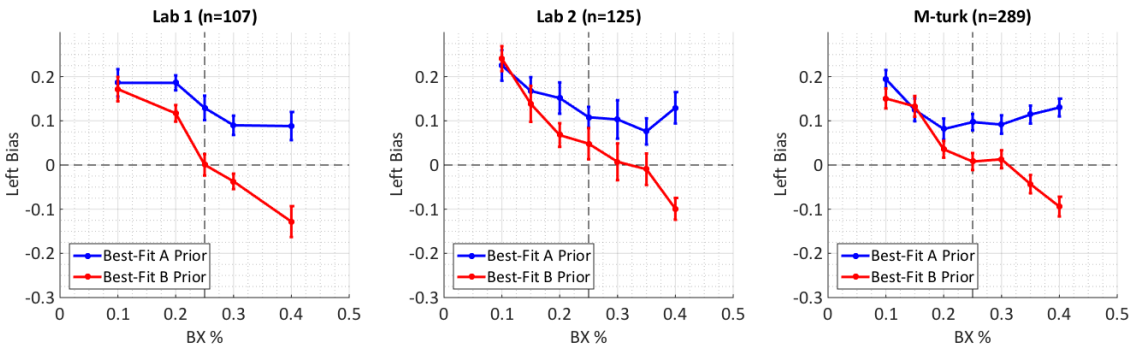


Figure 17 Best-fit A and B priors across three participant groups. The drift diffusion model was fit to each participant's data separately; the free parameters were A and B starting points, A and B thresholds, and a single drift rate, cue noise and T0 for all trial types. Lab 1 participants received 480 trials, while Lab 2 and M-turk participants received only 240 trials. The first 80 trials were discarded from the fits in all groups, as participants' behavior generally stabilized during those first two blocks. Across all participant groups, it appears that the A Prior decreases as a function of BX frequency, despite the fact that AX:AY frequencies remained constant. We quantify these effects in **Figure 18**.

Raw reaction time (RT) and accuracy data

For completeness, we also display how reaction times (RTs) and accuracy after A and B cues vary as a function of condition in **Figure 18**. However, since RT and accuracy sometimes exhibit different trajectories, we performed all subsequent analyses on the best-fit priors derived from our DDM fits, since these parameters take into account both RT and accuracy data, ensuring that any speed-accuracy trade-offs are taken into account.

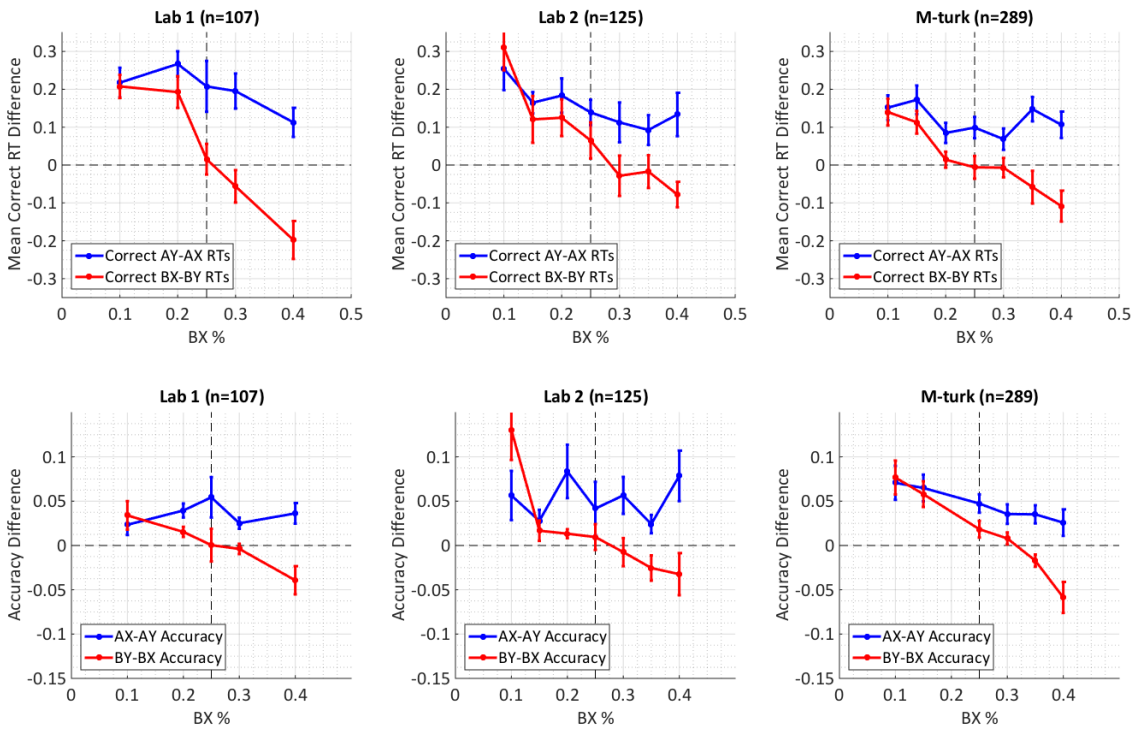


Figure 18 Mean reaction time and accuracy differences between left and right responses after A and B cues. The top panels show mean reaction time differences between correct right responses (AY and BX) and correct left responses (AX and BY) after A and B cues. The A trace (blue) is generally positive because right responses are generally slower than left responses, since the A cue always predicts the left response. On the other hand, the B trace (red) goes from positive to negative because the B cue varies from predicting the left response to predicting the right response. The bottom panels show analogous plot for accuracy. For each participant, the difference between left and right RTs or left and right accuracies was normalized (divided) by that participant's mean RT or mean accuracy, such that the values on the y-axis reflect proportions of the mean RT and accuracy.

Regression: Quantifying the effect of BX/BY frequencies on best-fit A prior

We tested the qualitative predictions of the three agents using a regression model that quantifies how the best-fit A prior varies as a function of experimental condition, i.e., the ratio of BY:BX trials.

Recall that Agent 1 predicted the A prior should not vary as a function of BY frequency, since it treats A and B priors independently (**Figure 16**, left panel). Agent 2 predicted that the A prior should decrease as BY frequency decreases, but level out (cease to decrease) as the two cues make increasingly different predictions (**Figure 16**, middle panel). Agent 3 predicted that the A prior should decrease with BY frequency when the two cues make similar predictions, but increase with BY frequency when they make different predictions (**Figure 16**, right panel).

We approximated these three curves using one intercept and two regressors. For Agent 1, a simple intercept term should be able to account for the data, since no slope is needed if A remains constant as a function of B. For Agents 2 and 3, we created two regressors, in which the influence of BY frequency on the A prior interacted with the similarity between A and B predictions. This is because both Agents 2 and 3 predicted two modes of A: first a decrease, and then either a leveling out or an increase. Thus, for both regressors, BY frequency ("BY%") was multiplied by the similarity between A and B predictions. The similarity was defined as:

$$ABsimilarity = |AX\% - BX\%|$$ [3]

While both regressors involved an interaction between BY% and ABsimilarity, we were able to roughly approximate the leveling out of Agent 2 by z-scoring the ABsimilarity term:

$$Regressor\ 1 = BY\% * zscore(ABsimilarity)$$

[3] The frequency of AX was always 40%, so the A cue always predicts the left response. When the frequency of BX trials is lowest (10%), the B cue also predicts the left response; therefore, the ABsimilarity term is highest when A and B response probabilities are most similar. As the frequency of BX trials increases to 40%, the B cue predicts the right response more strongly; therefore, the ABsimilarity term goes to 0 as A and B response probabilities diverge.

Similarly, we approximated the U-shaped curve of Agent 3 by z-scoring both the BY% and the ABsimilarity terms:

$$Regressor\ 2 = \ zscore(BY\%) * zscore(ABsimilarity)$$

Figure 19 shows how the two regressors vary as a function of experimental condition. Fitting this regression model to the best-fit A priors from all participants (from all three populations) revealed that the first regressor was significant (95% CI=[0.08 0.26]), while the second regressor was not (95% CI=[-0.05 0.13]), suggesting that Agent 2 provided the best explanation of participants' *implicit* priors (as measured by the best-fit starting point / threshold ratio). In other words, the fact that the A prior levels out as B gets dissimilar, but does not rise back up, suggests that even when participants adopt different priors for A and B in order to maximize reward rate, the frequencies of BY:BX continue to exert an influence on participants' A prior.

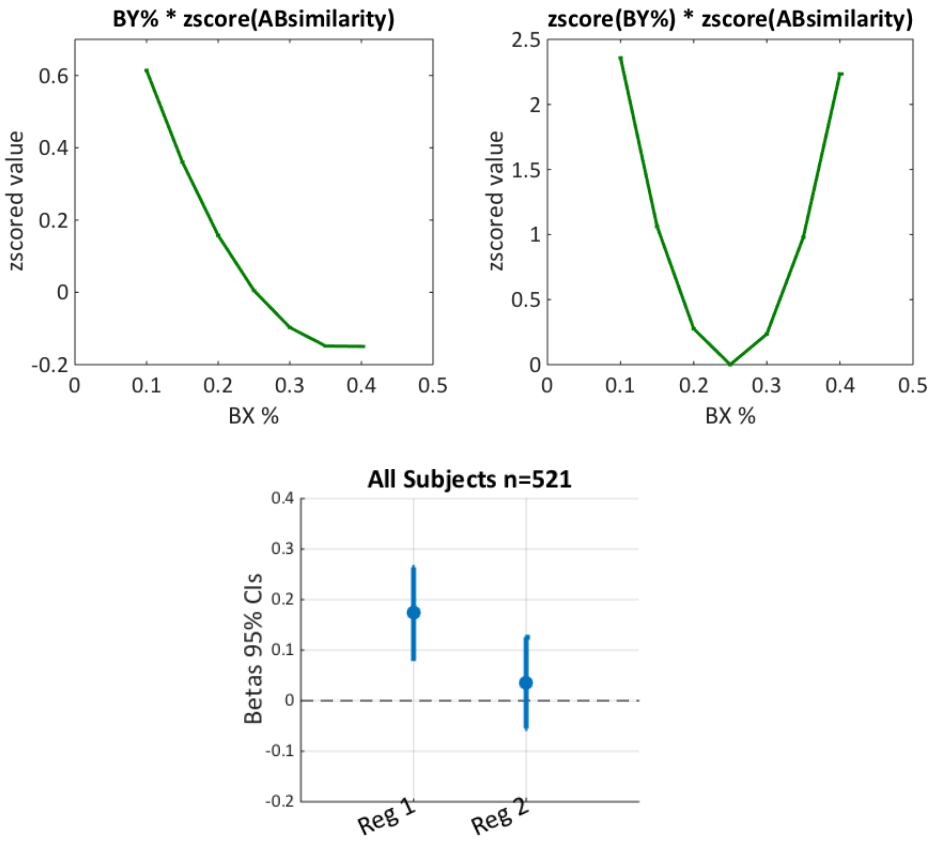


Figure 19 Best-fit A priors were significantly predicted by a regressor that approximates the strategy of Agent 2. The first regressor (left panel) approximates the curve predicted by Agent 2, in which the best-fit A prior should decrease with BY frequency and then level out when the cues make sufficiently different predictions. The second regressor (right panel) approximates the curve predicted by Agent 3, in which the best-fit A prior should decrease with BY frequency when the cues make similar predictions, but increase with BY frequency when the cues make different predictions. The bottom panel shows the 95% confidence intervals for the regression coefficients. The confidence intervals for Regressor 1 did not include 0, suggesting that participants' best-fit A priors were consistent with the predictions of Agent 2. In other words, the A prior continues to be influenced by the frequency of BY even when the two cues make different predictions (even when participants' best-fit A and B priors separate).

Explicit Response Priors

In addition to analyzing implicit response priors using reaction time and accuracy data on left and right responses, we also tested our hypothesis using participants' explicit estimates of the trial frequencies at the end of the experiment. During debriefing, participants were asked to

estimate the frequency of AX, AY, BX and BY trials. They were asked to provide a percentage for each trial type, such that the four numbers would sum to 100. **Figure 20** shows the ratio of AX:AY trials and BY:BX trials reported by our three participants groups. Note that we have smaller numbers of participants in all three groups relative to the implicit measure, since providing frequency estimates at debriefing was voluntary and some participants elected not to provide frequency estimates.

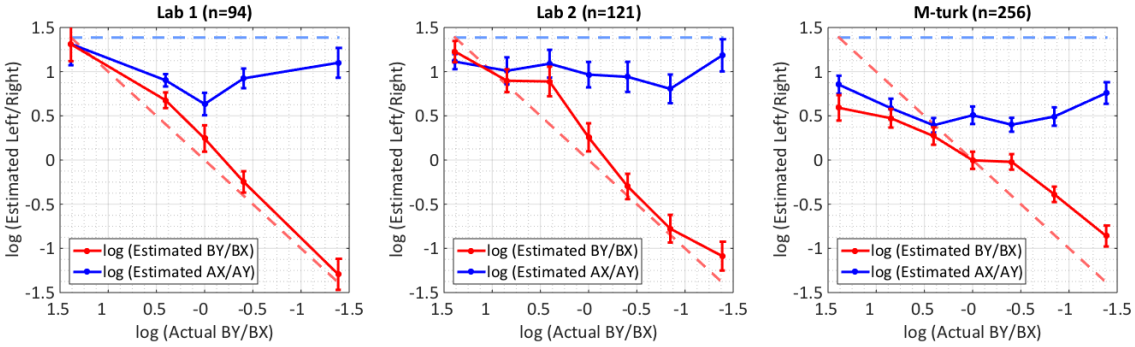


Figure 20 Reported AX:AY ratios and BY:BX ratios across three participants groups. The dashed lines reflect the true frequencies at which stimuli were presented, while solid lines reflect estimated frequencies reported by the participants in debriefing. The AX:AY ratio reflects the expected ratio of left:right trials after A cues, while the BY:BX ratio reflects the expected ratio of left:right trials after B cues. In the Lab 1 and M-turk groups, the estimated A ratio (blue trace) appeared to follow a U-shaped curve as a function of B, as predicted by Agent 3, which groups response probabilities across cues when they are similar enough. Moreover, the estimated B ratio (red trace) appeared to deviate from the true ratio and be pulled closer to A when the cues made similar predictions. We quantified these effects using a regression model displayed in **Figure 21**.

Regression: Quantifying the effect of BX/BY frequencies on reported AX:AY ratio

Just as for our implicit measure, we attempted to adjudicate between three different strategies: 1) never grouping A and B response probabilities; 2) always grouping A and B response probabilities; and 3) grouping A and B response probabilities only when the two cues make similar predictions about the response. We used the same regression model described above, but this time, to predict how the explicitly estimated AX:AY ratio varies as a function of BY frequency.

Unlike the implicit measure, fitting the regression model revealed that the second regressor (U-shaped curve of A as a function of B, reflecting strategy of Agent 3) significantly predicted explicitly estimated A ratios (95% CIs=[0.08 0.27]), while the first regressor (reflecting Agent 2) was not significant (95% CI=[-0.14 0.05]). This suggests that – at least for the explicit measure – response probabilities are grouped across cues only to the degree that they make similar predictions (**Figure 21**).

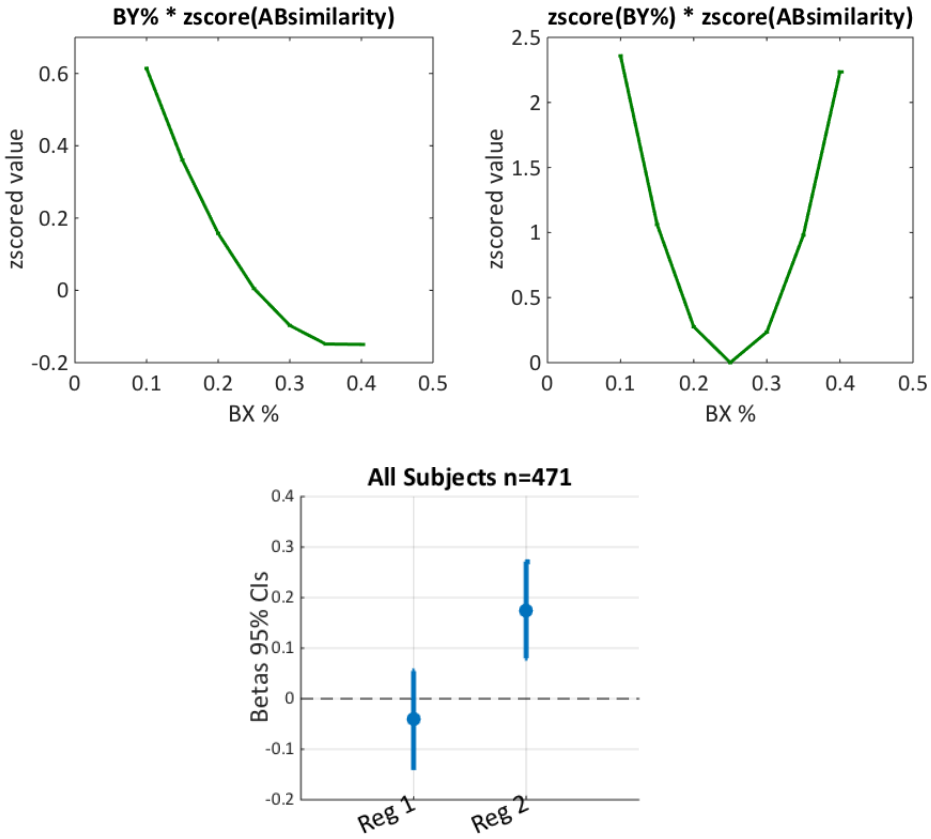


Figure 21 Explicitly reported A ratios were significantly predicted by a regressor that approximates the strategy of Agent 3. Just as in Figure 6, the first regressor (top left panel) approximates the curve predicted by Agent 2, while the second regressor (top right panel) approximates the curve predicted by Agent 3. The bottom panel shows the 95% confidence intervals for the regression coefficients. The confidence intervals for Regressor 2 did not include 0, suggesting that participants' reported Left:Right ratios for A cues followed a U-shaped curve as a function of B, consistent with the predictions of Agent 3.

Correlation between A and B Priors Varies by Similarity Between Cue Predictions

Thus far, we have shown that different measures of response priors implicate different strategies for how participants combined probabilities across cues. Participants' accuracy and

reaction time behavior on left vs. right responses after A vs. B cues ("implicit prior") show that the combination of response probabilities across cues never ceases (in line with Agent 2), even when A and B predict completely opposite responses. In contrast, subjective frequency estimates reported at debriefing ("explicit prior"), show that the A response prior (AX/AY frequency) is pulled towards the B response prior (BY/BX frequency) only when the two cues predict similar response probabilities. When the response probabilities after A and B cues become sufficiently different, the A prior seems to be pushed away from B and rises back to its original value (exhibiting a "U-shaped" curve in line with Agent 3).

The differences in grouping of cue predictions implied by the two measures suggest at least two possibilities for how participants formed and used representations of probability information relevant to the task. On the one hand, it could be that participants formed and used one set of representations while performing the task, and a different set for explicit report at the time of the debriefing. On the other hand, it could be that in both cases they used the same set of representations, which were probed at different times (progressively during training, and then in their final state at debriefing). To address these questions, we first performed additional analyses (described in the remainder of this section) and then an additional experiment (described in the next major section).

Implicit (best-fit) A and B priors remain positively correlated regardless of experimental condition

The regression analyses performed above (**Figure 19** and **Figure 21**) quantified how participants' implicit and explicit priors after A cues changed as a function of BX frequency.

A different way to gain insight into how A and B response priors are related is to examine, within each experimental condition, whether individual variability in the best-fit A prior correlates with individual variability in the best-fit B prior. **Figure 9** shows, for each condition, how the best-fit A prior varies with the best-fit B prior across participants. The strong positive correlation between them (especially for conditions in which BX% is low; that is, when the left response is more common for both cues) suggests that participants with a left bias after A cues also tend to have a left bias after B cues.

Of course, this correlation does not demonstrate that participants have "merged" the two priors internally. These parameters could be correlated simply because participants differ in the degree to which they are faster on the left than the right button, such that participants who are more left-biased overall would show greater left bias both on A and B trials. However, it is important to note that this relationship grows weaker for conditions when BX frequency is higher (when B predicts the right response more, while A predicts the left). For instance, when BX=35% (**Figure 22**, 6th panel), the negative correlation between these parameters implies that A and B response priors are no longer influenced by a common motor bias.

To test whether the correlation between A and B priors changes significantly as a function of experimental condition, we pooled the best-fit priors from all of our participants (including lab and M-turk participants) into a single regression, in which we tested whether the best-fit A prior could be predicted from the best-fit B prior, and whether this correlation was modulated by experimental condition (BX frequency):

$$A\ Prior \sim zscore(B\ Prior) + zscore(B\ Prior) * zscore(BX\ \%)$$

We found that both regressors had confidence intervals that did not include 0 (**Figure 22**, bottom panel). The fact that the interaction (between B prior and BX %) had a significantly negative coefficient (95% CI=[-0.26, -0.10]) shows that the correlation between the best-fit B Prior and the best-fit A prior was significantly lower when BX frequency was high (when A and B predict opposite responses), confirming the intuition we derived from the correlation plots. The fact that the B prior regressor was also significant (95% CI=[0.29, 0.45]) suggests that the strong positive correlation between these parameters nevertheless persists across all experimental conditions (otherwise, only the interaction would be significant). This confirms our finding above (**Figure 19**) that Agent 2 provided a better explanation of people's implicit priors, since it shows that the two parameters remain grouped within participants (a participant with a high A prior also tends to have a high B prior) even in conditions where BX% > 25%, when people have adopted different response priors for A and B. Thus, even though the plots in **Figure 17** show that the values of best-fit A and B priors have diverged from each other when the two cues make different predictions about the response, the fact that they remain correlated suggests some pooling of the implicit priors across cues, regardless of experimental condition.

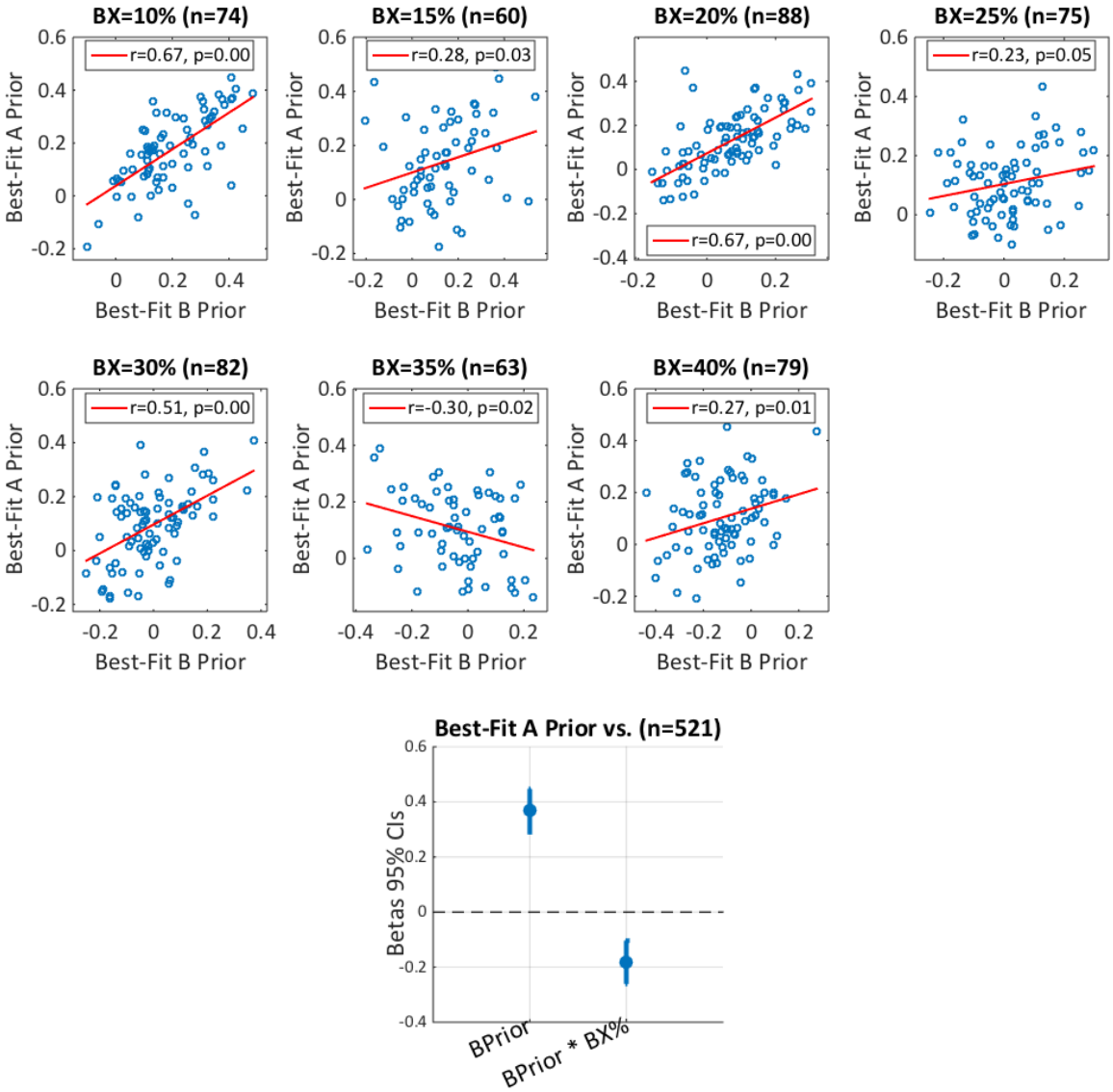


Figure 22 The correlation between the best-fit A and B priors significantly weakens as the cues make more different predictions. This analysis captures individual differences in the best-fit prior parameters for A and B cues. The significant positive coefficient for the "B Prior" regressor suggests that A and B priors are positively correlated across all experimental conditions, i.e., when a participant is left-biased after A, they tend to be left-biased after B as well. However, this correlation is significantly weaker when BX frequency is high, as evidenced by the significantly negative coefficient for the interaction between B prior and BX %. This means that the relationship between A and B priors weakens when B predicts very different response probabilities (right > left) from A. These results are consistent with our finding above that participants' implicit A and B priors are pulled towards each other, even when the two cues make very different predictions and even when their best-fit values are different (as predicted by Agent 2).

Explicitly reported A and B priors become anti-correlated when the two cues make opposite response predictions

We performed the same analyses described above using participants' explicit priors, to test whether the reported AX:AY ratio correlated with the reported BY:BX ratio. Such correlations are indicative of the way participants group trial types in each condition. For instance, if the reported AX:AY ratio correlates positively with the reported BY:BX ratio, it might mean that participants assume that if AX is frequent, then *BY* must be frequent as well, or vice versa. However, if the correlation is negative, it might mean participants assume that if AX is frequent, then *BX* must be frequent as well. More importantly, if these correlations change across experimental conditions, it would suggest that the frequencies participants experienced changed the way in which they grouped the cue-probe pairs (AX, AY, BX, BY).

The correlations between A and B ratios differed more across conditions in the explicit reports than in the implicit priors. **Figure 23** shows that the correlation between the reported AX:AY and BY:BX ratios varied from strongly positive when BX frequency was low (BX=10% right, BY=40% left), to zero when B did not make any predictions about the response (BX=25% right, BY=25% left), to strongly negative when BX frequency was high (BX=40% right, BY=10% left).

Again, we pooled all the reported ratios into a single regression model, testing whether the reported AX:AY ratio was significantly predicted by the reported BY:BX ratio, and whether this correlation was modulated by experimental condition (BX frequency):

$$A\ Ratio \sim zscore(B\ Ratio) + zscore(B\ Ratio) * zscore(BX\ \%)$$

In contrast to the results using implicit priors, the main effect of B ratio on A ratio was marginally above 0 (95% CI=[0.01, 0.17]), while the interaction between B ratio and BX frequency was substantially below 0 (95% CI=[-0.61, -0.46], see **Figure 23**, bottom panel). This suggests that there was a strong and persistent correlation across participants between the reported AX:AY and BY:BX ratios, but that the *sign* of this correlation changed significantly as a function of BX frequency, as is visible from **Figure 23**.

These results lend support to the idea that participants learned to group the cues and probes differently depending on the frequencies they experienced. Participants who saw more AX and BY trials, and pressed the left button more after both cues, showed strongly positive correlations, such that if a subject reported that they experienced a high AX:AY ratio, they also tended to report that they experienced a high BY:BX ratio. In contrast, participants who saw more AY and BX trials, but pressed the left button more after A and the right button more after B, had strongly negative correlations between A and B ratios. Accordingly, if they reported a high AX:AY ratio, they tended to report a low BY:BX ratio.

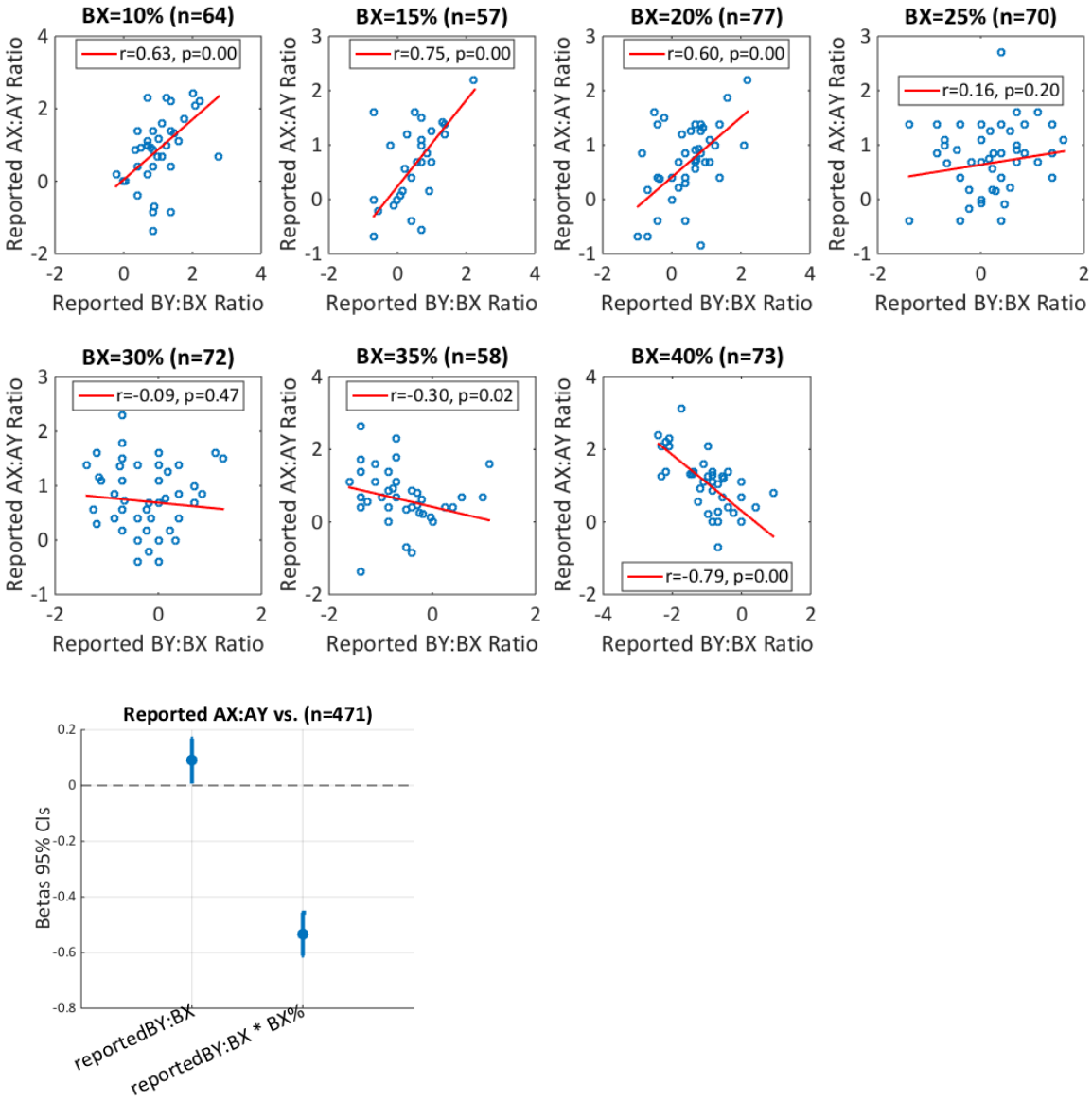


Figure 23 The correlation between reported AX:AY and BY:BX ratios flips from positive to negative as BX frequency increases. The significantly negative coefficient for the interaction between BY:BX ratios and the experimental condition (BX frequency) shows that the correlation between BY:BX and AX:AY ratios significantly decreases as BX % increases. In other words, when BX% is low (AX and BY are more frequent), participants who report high AX:AY ratios also tend to report high BY:BX ratios. On the other hand, when BX% is high (AX and BX are more frequent), participants who report high AX:AY ratios also tend to report low BY:BX ratios (which is the same as reporting high BX:BY ratios). This effect could be due to individual differences in how sensitive participants are to probabilities in general, and may only reflect the experimental condition that participants were given, or it may reflect something deeper about how participants learned to group their internal representations. The follow-up experiment will shed more light on this question.

Interpreting across-subject correlations between explicit A and B priors

These results may not seem surprising: participants grouped the stimulus pairs in precisely the way that the experimental condition grouped them, and the correlations between the reported ratios do not necessarily mean that participants' internal representations were grouped. For instance, one uninteresting reason why reported A and B ratios might be correlated is that participants differ in how sensitive they are to probability ratios. If a participant tends to underestimate probability ratios, then they might report smaller ratios for both AX:AY and BY:BX trials when AX and BY are more frequent. Similarly, in the condition where AX and BX are more frequent, such a participant would report lower AX:AY and BX:BY ratios, resulting in negative correlations when BX=40%. In other words, a participant could be learning A probabilities and B probabilities completely independently, but still report similar values for A and B ratios, because that participant underestimates *all* probability ratios to the same degree. Such an individual difference in probability underestimation or overestimation could potentially explain the difference in correlations we see across conditions. However, such an explanation could *not* account for the way the shape of the reported A ratio changes non-monotonically as a function of BX frequency (shown in **Figure 20** and **Figure 21**). The initial decrease and then increase of A as a function of B shown above implies an interaction between reported A and B probabilities. Since we know from the above analysis that reported A and B ratios interact in a non-monotonic way, it is at least possible that the change in correlations across participants is also a reflection of this interaction.

To verify this, we ran a follow-up experiment in which the frequencies of BX and BY trials changed mid-way through the study. If the cross-participant correlations between A and B

ratios truly reflect the way in which participants learned to group them internally, then changing the trial frequencies after they had learned these groupings might reveal transfer of the representations learned in the first part of the experiment to the second part. In other words, we posited that participants might continue to group stimulus pairs in the same way that they learned initially, at least for some period of time after the external environment had changed and no longer favored such grouping.

Experiment 2: Changing Trial Frequencies Within Participants

In Experiment 2, we sought to verify whether any representations that participants had learned in the first part of the study would transfer when we changed the trial frequencies within participants. Half of our participants (the MERGE group) received the BX=10% frequency condition for the first 5 blocks (200 trials) of the task, while the other half (the SPLIT group) received the BX=40% frequency condition for the first 5 blocks (**Figure 24**). In the second half of the experiment, the trial frequencies changed without notice to the BX=25% frequency condition, such that both groups received exactly the same frequencies in the last 5 blocks (200 trials) of the study. The transition was not signaled in any way. Participants were allowed to take a break (up to 2 minutes) after the end of each block before continuing to the next block, and the change happened during the break between the 5^{th} and 6^{th} blocks, making it particularly seamless. As in the previous experiments, participants were not told that the stimulus pairs had unequal frequencies and were not aware that the frequencies could change.

CHANGING FREQUENCIES WITHIN PARTICIPANTS

Merge Group

		Part 1	Part 2
AX	**Left**	**40%**	40%
AY	Right	10%	10%
BX	Right	10%	25%
BY	**Left**	**40%**	25%

Split Group

		Part 1	Part 2
AX	**Left**	**40%**	40%
AY	Right	10%	10%
BX	**Right**	**40%**	25%
BY	Left	10%	25%

Figure 24 Experimental design for Experiment 2. Participants were randomly assigned either to the MERGE or the SPLIT group. In the first 5 blocks of the experiment (Part 1), MERGE participants received the BX=10% condition, while SPLIT participants received the BX=40% condition. In the last 5 blocks (Part 2), the trial frequencies changed without notice to the BX=25% condition for both groups. We hypothesized that participants would learn to group response priors differently across A and B cues in the two conditions, and that this way of grouping would transfer to Part 2 of the experiment. Specifically, we tested whether the SPLIT participants would learn "separate" or "opposite" response priors for A and B cues, as suggested by the explicit measure, or whether they would learn response priors that were still grouped across cues, but more weakly than in the MERGE participants, as suggested by the implicit measure.

Hypotheses for Experiment 2

Our hypotheses for what might happen when the frequencies changed depended on whether we examined the implicit or the explicit reports. Remember that the implicit priors for A and B trials appeared correlated with each other regardless of experimental condition, only this correlation was weaker in the SPLIT than in the MERGE condition (**Figure 22**). Moreover, on aggregate, implicit A priors seemed to decrease as BX frequency increased and then level out

(**Figure 19**), consistent with the behavior of Agent 2, which continues to group priors across cues even as their predictions grow more dissimilar. If participants' behavior in Experiment 2 is in line with these findings from the implicit measure, then we would expect that the SPLIT participants would learn to group A and B response priors, though less strongly than the MERGE group.

On the other hand, examining participants' explicit reports provided us with very different predictions for Experiment 2. The U-shaped curve of A ratios as a function of B (**Figure 20**) was in line with Agent 3, which initially combines response priors across cues when they are similar, but then learns to separate A from B when the cues make different predictions. This would suggest that participants might learn to combine A and B priors in the MERGE group, but learn independent priors in the SPLIT group. However, the negative correlations between reported A and B ratios in the SPLIT condition suggest that the picture might be even more complicated than that: rather than learning entirely independent priors for A and B cues in the SPLIT condition, participants might instead learn that the priors are anti-correlated, such that they predict opposite responses but to the same degree. If this is the case, then both the MERGE and SPLIT groups would have learned to group A and B priors, only in opposite ways.

To summarize, while all the above analyses suggested that MERGE participants should learn to combine priors across cues, there were three possibilities for the SPLIT participants: 1) they might learn to combine priors, but more weakly than MERGE participants, 2) they might learn independent response priors for A and B, or 3) they might learn opposite response priors for A and B.

We tested which of these scenarios was more likely using three behavioral indices: 1) how the best-fit A prior changes as a function of B frequencies, 2) across-participant correlations between A and B priors, and 3) within-participant correlations between A and B priors across blocks.

Results of Experiment 2

Changing trial frequencies changed participants' behavior

To test how the best-fit A and B priors changed within participants across blocks, we fit our drift diffusion model to each block (40 trials) separately, resulting in best-fit parameter values for each of the 10 blocks. **Figure 25** shows how the best-fit A and B priors evolved over time, depending on whether a participant had been assigned to the MERGE or the SPLIT group.

To verify whether participants changed their priors in response to the change in trial frequencies, we tested whether the difference between A and B priors for each participant changed significantly between Part 1 and Part 2 of the experiment. Since participants appeared to adapt during the first two blocks and did not stabilize until the third block (**Figure 25**), we tested for a significant difference between blocks 3-5 and blocks 8-10, discarding the first two blocks after each "transition" during which participants might still be adapting to the change. The differences in priors for each participant were averaged across blocks 3-5 and blocks 8-10, obtaining a single value for Part 1 and a single value for Part 2 of the experiment in every participant.

We found that the mean difference between A and B priors for MERGE participants increased from almost 0 in blocks 3-5 (M=-0.004, SD=0.15) to 0.08 in blocks 8-10 (M=0.08,

SD=0.16), suggesting that MERGE participants were sensitive to the fact that response frequencies for A and B diverged in Part 2 of the experiment (*p*=0.009, *z-value*=2.6, Wilcoxon signed rank test). Similarly, for SPLIT participants, the mean difference between A and B priors decreased from 0.29 in blocks 3-5 (*M*=0.29, *SD*=0.24) to 0.21 in blocks 8-10 (*M*=0.21, *SD*=0.28), suggesting that SPLIT participants were also sensitive to the fact that response frequencies for A and B became closer together in Part 2 (*p*=0.017, *z-value*=2.4).

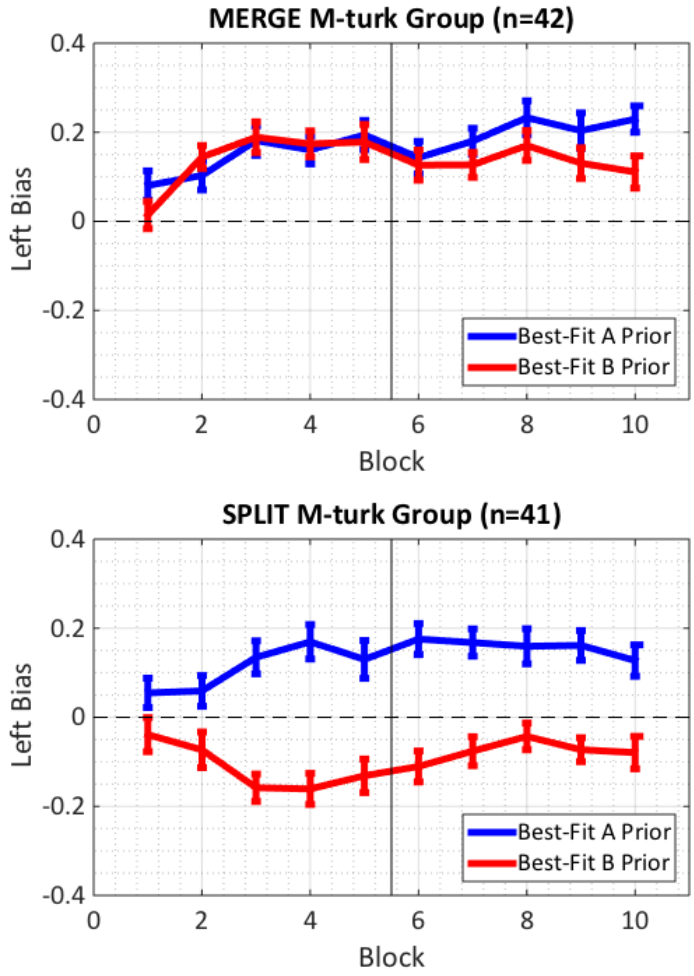


Figure 25 Best-fit A and B priors across blocks, averaged across MERGE and SPLIT participants. The drift diffusion model was fit separately for every block in every participant, in order to track how the prior (starting point / threshold) parameter changed across blocks. Each block contained 40 trials. Error bars reflect standard errors of the mean across participants. The black vertical line between blocks 5 and 6 reflects the time at which trial frequencies changed for both groups to the BX=25% condition. We found that the difference between A and B priors changed significantly between blocks 3-5 and blocks 8-10, increasing significantly for the MERGE group, while decreasing significantly for the SPLIT group, suggesting that participants had changed their behavior in response to the change in trial frequencies.

Having obtained confirmation that participants adapted their behavior after the change in trial frequencies, we could now verify whether the way they grouped response priors across cues nevertheless transferred from Part 1 to Part 2.

Best-fit A prior as a function of BX frequency

We next verified how the effect of BX frequency on the A prior differed between SPLIT and MERGE participants. If both participant groups learned to combine A and B response priors, then varying the frequency of BX trials should change the best-fit A prior. In Part 1 of the experiment, the SPLIT group experiences fewer left responses overall than the MERGE group, which should decrease the left bias on A for the SPLIT group relative to the MERGE group. On the other hand, if SPLIT participants learn independent or opposite response priors for both cues, then their A prior in Part 1 should not be pulled down towards B, and should be as high as the A prior for MERGE participants. We found that the best-fit A prior in blocks 3-5 for MERGE participants (M=0.18, SD=0.18) could not be distinguished from the best-fit A prior in blocks 3-5 for SPLIT participants (M=0.14, SD=0.21), suggesting that the A prior did not seem to be significantly pulled down towards B in the SPLIT group (p=0.63, z-*value*=0.48, two-sided Wilcoxon rank sum test). Although we cannot prove the null hypothesis that the A prior is identical in value between the two groups, this result favors the interpretation that SPLIT participants learned either independent or opposite priors for A and B in Part 1 of the experiment.

As for our predictions for Part 2, if both participant groups learned to combine A and B response priors, then for MERGE participants, the A prior should decrease (in left bias) from Part 1 to Part 2 of the experiment, since the overall left response frequency decreases. For SPLIT participants, the A prior should *increase* from Part 1 to Part 2, since the overall left frequency increases. On the other hand, if SPLIT participants learned independent priors for A and B cues, then for SPLIT participants, the A prior should not change as the B prior changes.

The third possibility is that SPLIT participants learned opposite response priors for A and B cues. In that case, it is possible that participants might maintain opposite response priors, such that A remains positive while B remains negative in Part 2.

Contrary to all our predictions for MERGE participants, we found that the best-fit A prior continued to increase in Part 2, despite the fact that overall left frequency decreased. Its mean value increased from 0.18 (*SD*=0.18) in blocks 3-5, to 0.22 (*SD*=0.18) in blocks 8-10. Although this increase was marginally significant (*p*=0.076, *z-value*=1.8, Wilcoxon signed rank test), it suggested that the best-fit A prior was not pulled down towards B as we had predicted. Examining the time course of the A prior in **Figure 25** shows that it continues to grow in left bias despite the change in B frequencies, while the B prior stops growing in left bias and slightly decreases. In fact, the B prior for the MERGE group decreased from *M*=0.18 (*SD*=0.18) in blocks 3-5 (Part 1) to *M*=0.14 (*SD*=0.17) in blocks 8-10 (Part 2), though this decrease was only marginally significant (*p*=0.09, *z-value*=−1.7). The fact that the B prior remains significantly left-biased for Part 2 ($p<10^{-4}$, *z-value*=4.0), even though B no longer predicts either the left or the right response (BX=25%, BY=25%) suggests that A might be exerting an influence on B, pulling it up away from 0, as expected if participants are combining response frequencies across cues. (The remaining left bias on B could also be due to carryover from Part 1. However, the fact that the relative difference between A and B priors increased significantly in response to the change in frequencies suggests that this pattern may not reflect a simple carryover of behavior from Part 1. In fact, if B had continued in its previous trajectory, it might have continued to increase together with the A prior, but instead the A and B priors diverged in blocks 7-10.) Regardless,

the continuous increase in A Prior is puzzling, and suggests a "building up" of the left bias on A over time that we did not expect.

For SPLIT participants, the best-fit A prior remained the same between blocks 3-5 (M=0.14, SD=0.21) and blocks 8-10 (M=0.15, SD=0.19; p=0.88, z-value=0.15), whereas the best-fit B prior significantly increased in response to the change in B response frequencies, increasing from M=−0.15 (SD=0.18) in blocks 3-5 to M=−0.06 (SD=0.17) in blocks 8-10 (p=0.001, z-$value$=3.2). These findings are perfectly consistent with the interpretation that SPLIT participants learned independent priors for A and B, since the B prior adapted to the change in B frequencies, while the A prior did not. The fact that the B prior remained significantly right-biased in blocks 8-10 (p=0.034, z-value=−2.1) hints that participants may have learned that B has an opposite bias from A, but this right bias might also simply be due to inertia (carry-over from Part 1) and cannot be interpreted as such.

To summarize, SPLIT participants' priors in both Part 1 and Part 2 of the experiment were consistent with the interpretation that they learn separate response priors for A and B. In Part 1, the A prior did not appear to be pulled down towards B in SPLIT relative to MERGE participants. In Part 2, the B prior adapted to the change in BX/BY frequencies, while the A prior did not. However, MERGE participants' behavior did not unfold as we had predicted: rather than decreasing their left bias on A in response to the decrease in overall left frequencies, the left bias on A continued to increase, as if participants' left bias on A built up over blocks. Their B prior did decrease slightly as B frequencies changed, but remained significantly left-biased, possibly because it was being pulled up towards A.

Across-participant correlations between best-fit A and B priors

We next examined whether individual variability in the best-fit B prior was correlated with individual variability in the best-fit A prior and whether this correlation differed between MERGE and SPLIT participants (**Figure 26**).

If SPLIT participants have learned to combine A and B response priors, only more weakly than MERGE participants, we might expect to see a positive correlation between the best-fit priors across participants that is weaker for SPLIT than for MERGE participants. On the other hand, if SPLIT participants have learned independent priors for A and B, the correlation between A and B priors for these participants might be indistinguishable from 0, or even negative if participants have learned opposite response priors for the two cues.

As above, best-fit priors were averaged across blocks 3-5 for Part 1 and across blocks 8-10 for Part 2. For MERGE participants, we found that the correlation between A and B priors across participants was strongly positive for both Part 1 (r=0.62, $p<10^{-4}$) and Part 2 (r=60, $p<10^{-4}$) of the experiment. In other words, participants who displayed strong left biases after A also tended to have strong left biases after B, even in Part 2, when B priors were significantly less left biased than A priors (p=0.003, *z-value*=3.0). This strong positive correlation could be due to the fact that MERGE participants developed an overall left bias over the course of the experiment. If the strength of the left bias differs across participants, it would result in a positive correlation between the best-fit A and B priors, since it would facilitate left responses after both cues. Thus, this strong positive correlation cannot as such be taken as a sign that participants are “merging” their representations of A and B priors.

For SPLIT participants, A and B priors were weakly correlated across participants in Part 1 (r=0.19, p=0.24) but this correlation did not reach significance. In Part 2, the cross-participant correlation became slightly negative, but was very far from reaching significance (r=−0.11, p=0.49), suggesting that a SPLIT participant's best-fit A prior could not be predicted by looking at their best-fit B prior. These results are again consistent with the interpretation that SPLIT participants developed independent priors for A and B cues.

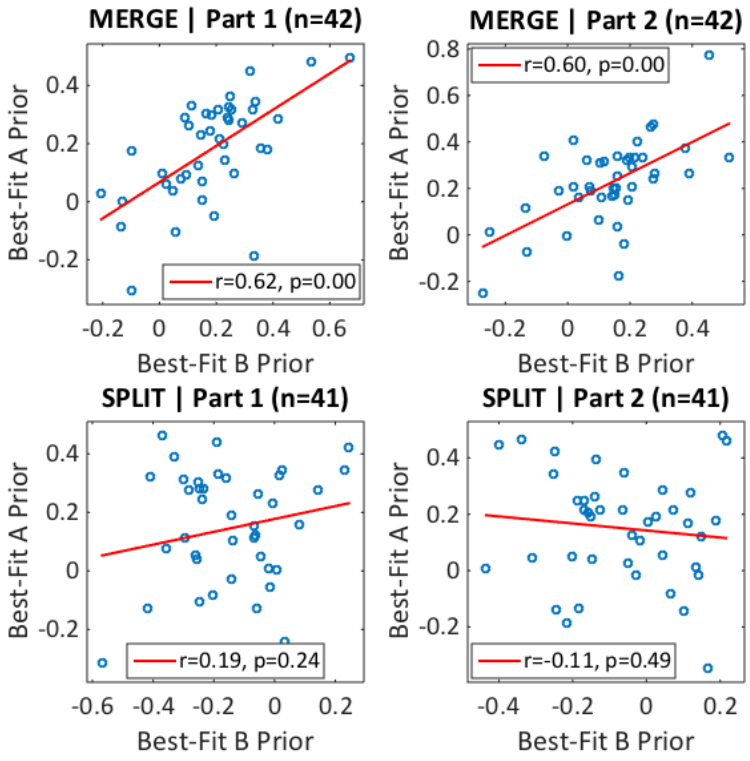


Figure 26 Across-participant correlations between best-fit A and B priors were significantly positive for MERGE participants but did not differ from 0 for SPLIT participants. In every participant, best-fit A and B priors were averaged across blocks 3-5 for Part 1 and across blocks 8-10 for Part 2. They were then correlated with each other across participants. The strong positive correlation between these parameters for MERGE participants may be due to an overall left bias that affects all trials, but that differs in intensity across participants. The fact that the correlation did not differ significantly from 0 for SPLIT participants suggests that these participants may have learned independent response priors for A and B cues.

Within-participant correlations between best-fit A and B priors

The greater number of trials in this experiment enables us to examine correlations between best-fit A and B priors *within* participants across blocks. To maximize the number of data points for a correlation (which was already quite small), we correlated the best-fit A and B priors within a participant across the first 5 blocks for Part 1 and the last five blocks for Part 2.

We first examined within-participant correlations in Part 1 of the experiment, when our two participants groups received different trial frequencies. **Figure 27** (left panel) shows the distribution of within-participant Pearson's correlation values across participants for the first five blocks. As is visible from this figure, the median correlation value for MERGE participants (*median*=0.33, *SD*=0.55) was higher than the median correlation value for SPLIT participants (*median*=−0.29, *SD*=0.47), though there was tremendous variability in the correlation values across participants within each group. This difference in medians was significant according to a Wilcoxon rank sum test ($p<10^{-4}$, *z-value*=4.0), and a Wilcoxon sign rank test showed that MERGE correlations were significantly skewed towards positive (*p*=0.005, *z-value*=2.8), while SPLIT correlations were significantly skewed towards negative (*p*=0.004, *z-value*=−2.9).

This result (the negative within-participant correlations between A and B priors in the SPLIT group) is in conflict with our findings about implicit priors in our previous experiment, in which we found that participants in the SPLIT condition displayed weakly *positive* correlations between A and B priors across participants (**Figure 22**). Note that this result is even in conflict with the across-participant correlations in the same dataset: above, we found a weakly positive, insignificant correlation between A and B priors for SPLIT participants during Part 1 of the experiment (**Figure 26**). This suggests that within-participant correlations between the best-fit

priors may provide additional insight into how priors are being grouped within participants that could not be gained from correlating the parameters across participants. This result is, however, consistent with the explicitly reported priors in our previous experiment, where we found that AX:AY ratios correlated negatively BY:BX ratios across participants in the SPLIT condition. In interpreting that result, we cautioned that a negative correlation between A and B priors does not necessarily imply grouping of the internal representations of these parameters. In fact, a participant could independently learn that A is left-biased while B is right-biased, but since learning takes place across blocks (**Figure 25** shows that both A and B increase in response bias across blocks), these negative correlations could be caused simply by the process of *independently* learning the A and B probabilities provided by the task environment (in other words, the true probabilities of left after A and right after B are the same but opposite across cues, and these true probabilities might be the reason that participants' priors are correlated during learning).

To ensure that these correlations are not caused by learning the probabilities in the environment, we examined them during Part 2 (blocks 6 to 10) of the experiment, when both SPLIT and MERGE participants are receiving precisely the same trial frequencies. **Figure 27** (right panel) shows the distribution of within-participant correlations for the last 5 blocks. Numerically, the median correlation value for the MERGE group remained positive (median=0.12, SD=0.47), while that of the SPLIT group remained negative (median=−0.11, SD=0.46), although this difference between groups was only marginally significant according to a Wilcoxon rank sum test (p=0.07, *z-value*=1.8). This suggested that some of the way in which priors were grouped across cues may have transferred to Part 2 of the experiment, even though

the true response frequencies were now totally dissociated between A and B (A predicted the left response, while B was neutral).

Since participants may have adapted to the new trial frequencies by the end of Part 2 of the experiment, and possibly changed the way in which they grouped stimuli, we additionally verified whether the first blocks after the change (blocks 6-9) still displayed within-participant correlations that resembled those in Part 1. When the last block of the experiment was excluded from the correlation, we found that the difference in correlation values between MERGE and SPLIT groups during blocks 6-9 grew much stronger (Figure 14, bottom right panel). Specifically, the median correlation for MERGE participants (median=0.18, SD=0.57) was significantly higher than the median correlation for SPLIT participants (median=−0.37, SD=0.51) during those blocks ($p=0.003$, z-value=−2.9).

Our results show that A and B priors in MERGE participants remained positively correlated in Part 2 of the experiment, when the environment no longer provided any correlations in the response frequencies between the two cues. Moreover, the two priors remained negatively correlated in SPLIT participants, suggesting that these participants may have learned that A and B predict opposite response probabilities, just as our explicit measure in the previous experiment had predicted. Since both groups are receiving exactly the same trial frequencies in Part 2, these differences in within-participant correlations must be due to learning that took place in Part 1. In other words, the trial frequencies that participants experienced during Part 1 influenced the way that grouped response probabilities across contexts, in a way that transferred even when the frequencies changed.

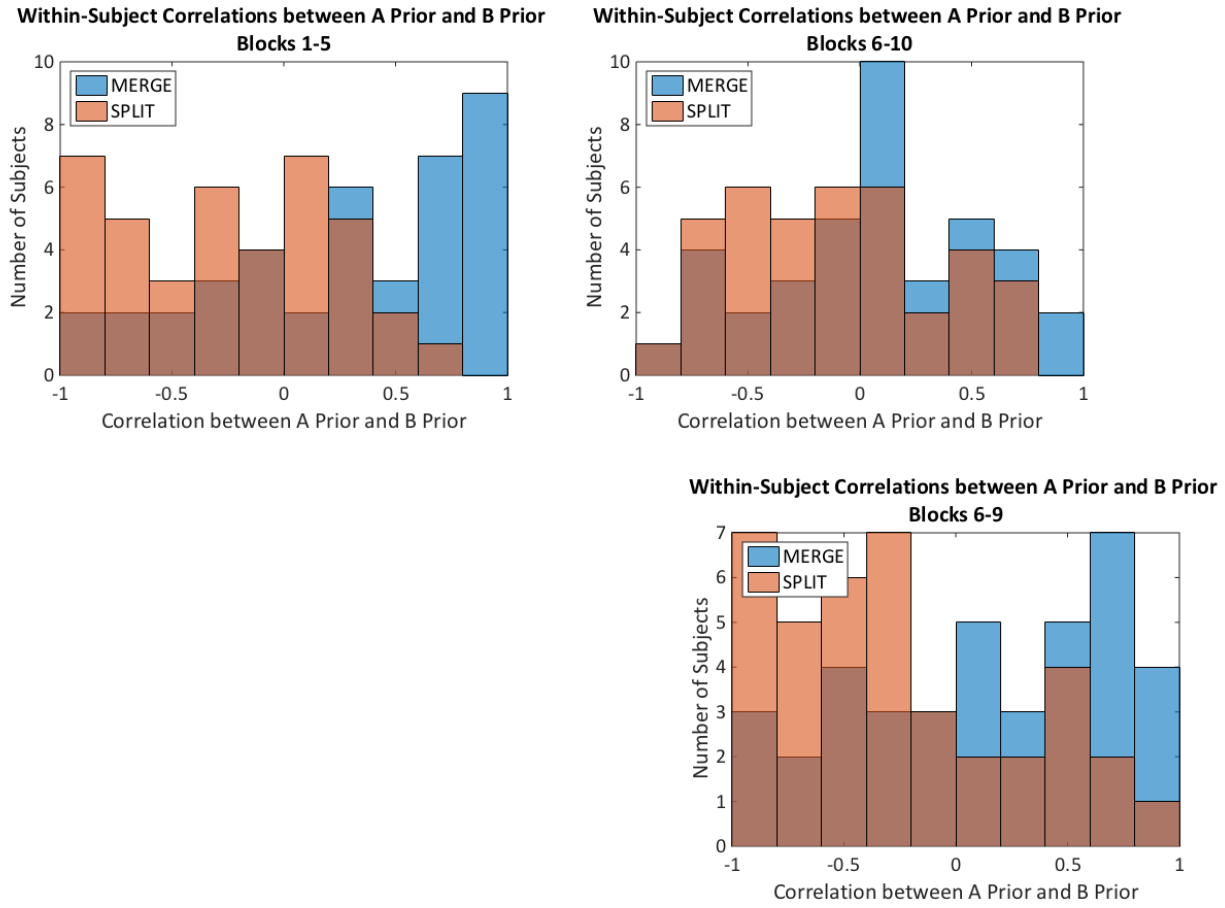


Figure 27 Distribution of within-participant correlations between best-fit A and B priors across blocks 1-5 and blocks 6-10 of Experiment 2. For each participant, the best-fit values of the two parameters were correlated across blocks (the model had been fit to each block of 40 trials separately). A and B priors tended to be positively correlated across blocks in MERGE participants, but negatively correlated in SPLIT participants. In blocks 1-5, when the two groups are receiving different trial frequencies, these correlations could be introduced by the environment: MERGE participants are experiencing the same response probabilities after A and B, while SPLIT participants are experiencing opposite response probabilities after A and B. However, these differences in within-participant correlations persist during blocks 6-9 (Part 2 of the experiment), when both participant groups are experiencing exactly the same trial frequencies (BX=25%). In the BX=25% condition, the B cue does not predict either left or right responses, such that A and B response frequencies are entirely dissociated. The fact that A and B priors remain tethered in MERGE and SPLIT participants, even when the environment is no longer affording any correlations, suggests that these correlations reflect true differences in how participants learned to group the priors internally.

Individual differences in within-participant correlations predict between-participant correlations

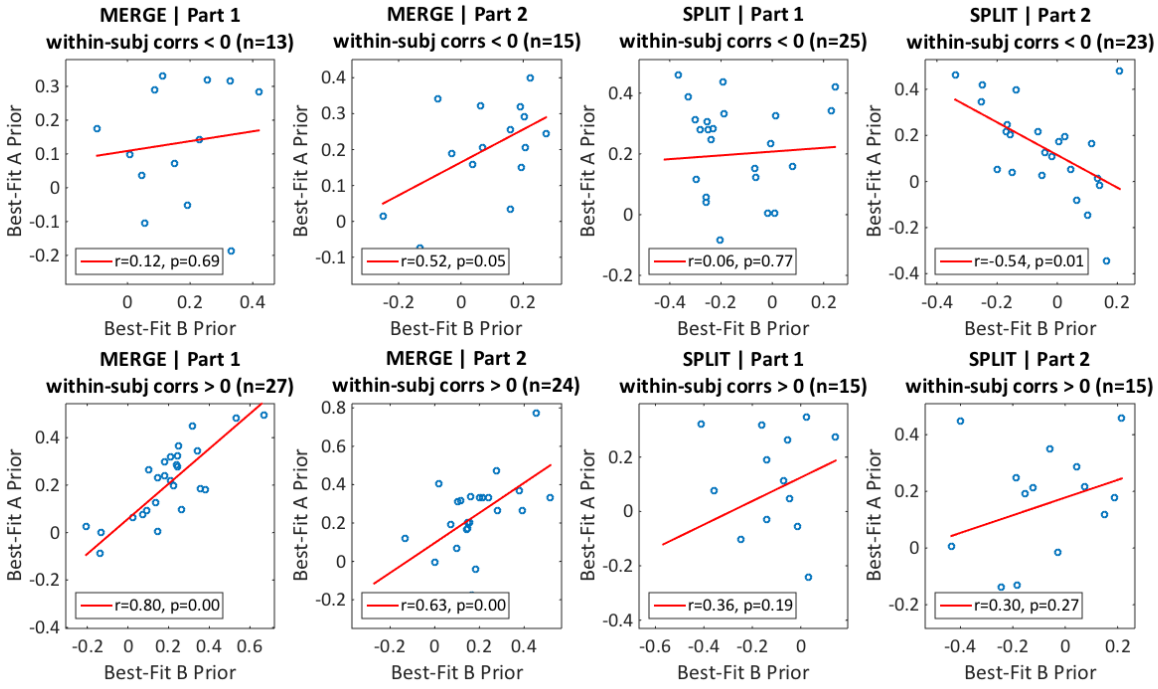
We next investigated why across-participant correlations in A and B priors might show different patterns from within-participant correlations. Recall that the across-participant correlations were strongly positive for MERGE participants (**Figure 26**, top panels), but indistinguishable from 0 for SPLIT participants. On the other hand, the two groups displayed "symmetric" within-participant correlations, weakly positive for the MERGE group and weakly negative for the SPLIT group.

In theory, the two measures should match: if A and B priors show negative correlations within participants, such that a rise in A is usually accompanied by a fall in B, then participants who ultimately develop strongly positive A priors should also develop strongly positive B priors. Such within-participant dependencies should generally create across-participant dependencies as well, so long as there is individual variability in the overall strength of priors across participants.

One possible explanation for the discrepancy could be that all participants vary in the degree to which they display overall response biases, preferring either the left or right response, after both A and B cues. Such response biases could increase between-participant correlations in both SPLIT and MERGE participants, since both A and B priors would be skewed to the left or to the right for a given participant. If such biases affect all blocks equally (adding some degree of left or right bias to the entire time course), they would not alter within-participant correlations, which are driven by variability across blocks.

To summarize, if the degree of response button preference varies across participants, it would increase the positive correlation between A and B priors across both SPLIT and MERGE participants. For MERGE participants, the within-participant correlations tend to be positive as well, such that the two effects might combine to produce an even stronger positive correlation than was present within participants. For SPLIT participants, since within-participant correlations tend to be negative, it is possible that the two effects might cancel each other out (any variance in overall response biases across participants might cancel out the effect of negative correlations between A and B priors *within* participants).

To characterize the relationship between the two types of correlations (within and across participants), we divided both SPLIT and MERGE participants into those who displayed positive correlations between A and B priors within participants and those who displayed negative correlations within participants. **Figure 28** shows the across-participant correlations between A and B priors for each of those subsets: the top panel displays participants with negative correlations across blocks, while the bottom panel displays participants with positive correlations across blocks. Regardless of condition (SPLIT / MERGE) or the phase of the experiment (Part 1 vs. Part 2), we found that between-participant correlations were more positive across participants who displayed positive *within-participant* correlations, than participants who displayed negative within-participant correlations.



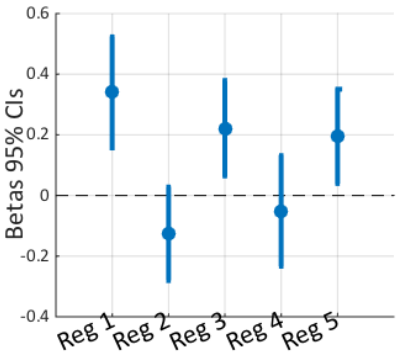


Figure 28 Across-participant correlations between A and B priors are modulated by within-participant correlations. For each group (SPLIT/MERGE) and each part of the experiment (blocks 1-5 vs. blocks 6-10), we recalculated the cross-participant correlation between A and B priors, separately for participants who displayed negative within-participant correlations between them (top row) and those who displayed positive ones (second row). The cross-participant correlations appeared stronger and more positive for participants who also showed positive correlations across blocks (second row). A regression model (bottom panel) showed that this effect was significant. Regressor 1 quantified the overall effect of B prior on A prior. Regressor 2 was the value of the within-participant correlation for a participant. Regressor 3 was the interaction between B prior and within-participant correlation. Regressor 4 was the group identity (SPLIT or MERGE). Regressor 5 was the interaction between group and B prior: this effect is significant because the correlation between A and B across participants is much stronger for MERGE than for SPLIT participants. The fact that Regressor 3 has confidence intervals above 0 shows that within-participant correlations significantly predict between-participant correlations. The fact that Regressor 1 is significant shows that A and B priors are correlated across participants overall, regardless of group identity.

To verify the significance of this effect, we fit a linear regression in which a participant's A prior (averaged across blocks) was predicted by their B prior (averaged across blocks), their within-participant correlation between the two priors, the interaction between the B prior and within-participant correlation, the group (SPLIT or MERGE), and the interaction between group and the B prior.

$$A\ Prior \sim B\ Prior + withinSubjectCorr + B\ Prior * withinSubjectCorr + \ group + B\ Prior * group$$

In other words, we tested whether within-participant correlations significantly predicted the correlation between B prior and A prior (third regressor). The interaction between B prior and group was a potential confound: we had already found that the correlation between A and B across participants was significantly stronger for MERGE than for SPLIT participants. Since MERGE participants also display more positive within-participant correlations, while SPLIT participants display more negative ones, we needed to ensure that any effect of within-participant correlations was not due simply to group membership.

As expected, we found a significant interaction between B prior and group (Regressor 5, **Figure 28**), reflecting the fact that MERGE participants showed stronger cross-participant correlations. The interaction between B prior and within-participant correlations (Regressor 3) was significant above and beyond this effect, suggesting that even within MERGE participants, groups of participants who showed negative correlations across blocks tended to have weaker across-participant correlations; similarly, within SPLIT participants, those who showed positive

correlations across blocks tended to have positive across-participant correlations, as we intuited from **Figure 28**.

These results are at least consistent with our interpretation that between-participant correlations are partially driven by within-participant correlations, but that additional factors produce an overall positive correlation between A and B priors across participants (Regressor 1), regardless of experimental condition. For instance, if participant-specific button preferences create a positive correlation between A and B priors across participants, this would cancel out the negative correlations within SPLIT participants, while boosting the positive correlations within MERGE participants.

If the above interpretation is correct, and between-participant correlations are confounded by variability in "handedness", then within-participant correlations may be a more sensitive measure of our mechanism of interest – whether A and B priors are tethered within participants.

Conclusion of Experiment 2

The results of Experiment 1 suggested that participants may have learned to combine response probabilities across contexts in different ways depending on the trial frequencies they experienced. They appeared to group probabilities across cues when the two made similar predictions, and to separate them more when the cues predicted different response probabilities, although the precise pattern varied between implicit and explicit priors. While implicit priors appeared grouped regardless of similarity, explicit reports showed that when A and B predicted sufficiently different responses, A and B ratios were no longer positively

correlated, but instead became negatively correlated. The aim of Experiment 2 was to test whether any of the learning that took place transfers when the trial frequencies change, which would prove that participants learned "meta-parameters" of the task that changed the way in which new observations were organized, but additionally reveal the type of structure that participants had learned.

To do this, new M-turk participants were randomly assigned to either a MERGE group, in which A and B predicted the same response probabilities (80% left 20% right) or a SPLIT group, in which A predicted the left response more while B predicted the right response more. After 5 blocks of 40 trials, the response probabilities changed for the B cue to 50% left and 50% right for both groups, while the A probabilities remained unchanged. We tested whether participants would adapt to the change in B frequencies, and whether the A prior would also change in response to a change in the B prior, despite the fact that A probabilities had remained objectively the same.

We found that both groups adapted significantly to the change, separating their A and B priors after the change in the case of the MERGE group, and bringing their priors closer together in the SPLIT group. Despite this adaptation, we found, using both between-participant and within-participant correlations, that they had learned to group their observations differently across contexts depending on which trial frequencies they started with. MERGE participants displayed positive correlations their A and B priors, such that if the A prior grew in left bias, the B prior would do so as well. SPLIT participants displayed negative correlations, such that on blocks when their A prior became more left biased, their B prior became more right-biased. These differences in within-participant correlations remained significantly

different after the trial frequencies had changed, even when both groups were receiving precisely the same trial frequencies objectively. This suggests that participants had learned a relationship between A and B response probabilities, which continued to impact their behavior even when those relationships were no longer present in the environment.

The fact that the SPLIT group learned negative correlations between A and B priors is in line with the patterns present in the explicit probability estimates in our previous experiment. Those explicit reports showed negative correlations between reported A and B response ratios in the SPLIT condition. On the other hand, the implicit reports simply showed weaker positive correlations between A and B priors across participants, rather than negative ones. Experiment 2 revealed that between-participant correlations in the SPLIT condition could appear null or even positive, despite the fact that the two parameters show negative correlations within participants. We speculated that positive correlations between participants could be driven by individual differences in "handedness" across participants (with some randomly preferring the left and some the right), creating a motor bias that would affect both A and B trials and therefore make them appear correlated across participants. On the other hand, the explicit reports were not affected by such confounds, and their correlations between them might therefore have given a cleaner index of people's internal grouping strategy.

Our results leave open to interpretation the meaning of a negative correlation between A and B response priors. We had mentioned in the Introduction that our task has the virtue of dissociating merging of predictions across cues from disregarding of cue information completely, since even when the two cues make similar response predictions, they are still necessary for making the appropriate response to a stimulus. Because of this feature that the

cue flips the correct response to the stimulus, the stimulus (X/Y) dimension is perfectly anti-correlated with the response dimension (L/R). In other words, when the two cues predict the same response, they in fact predict opposite stimuli, and when they predict different responses, they predict the same stimulus. It is therefore entirely possible that a negative correlation between A and B response priors is in fact caused by a positive correlation between A and B stimulus predictions, such that instead of having learned to average response probabilities across cues, participants may have learned to average stimulus probabilities across cues in the SPLIT condition.

Relating Explicit and Implicit Measures: Do distortions in probability estimates account for sub-optimality in decision-making?

Having explored how participants combined predictions across cues using both implicit and explicit priors, we next examined the relationship between the two measures. Specifically, we tested whether distortions in probability learning (as measured by their explicit reports of the probabilities) could account for sub-optimality in participants' implicit priors on the task. (All of the below analyses pertain to the data from our original experiment, and do not include data from Experiment 2.)

Participants' Explicit Estimates Predict Their Implicit Priors

According to the results from our original experiment, participants' implicit priors showed different patterns from their explicit reports, at least in the way their A priors were influenced

by BX/BY frequencies. Implicit A priors decreased as a function of BX/BY frequencies (consistent with Agent 2), while explicit A reports showed a U-shaped curve as a function of BX/BY frequencies (consistent with Agent 3). Despite these differences, M-turk participants in particular seemed to display similar patterns in both their implicit priors (**Figure 17**, right panel) and explicit reports (**Figure 20**, right panel).

To quantify the relationship between the implicit and explicit measures, we fit a regression verifying whether a participant's explicitly reported A ratio (AX/AY) could predict their implicit A prior (derived from fitting the DDM), even when accounting for the effect of experimental condition:

$$A\ Prior \sim Estimated\ A\ Ratio + BY\ \% * zscore(ABsimilarity)$$

The second regressor was identical to "Regressor 1" in the first two regression models (**Figure 19**, top left panel), which was shown to capture the effect of BY frequency on the implicit A prior. Fitting this regression to the best-fit A priors from all participants who provided explicit reports at debriefing (*n* = 471) revealed that the estimated A ratio significantly predicted the best-fit A prior (**Figure 29**, left panel).

We fit an analogous model to the best-fit B priors:

$$B\ Prior \sim Estimated\ B\ Ratio + BY\ \%$$

This time, the effect of experimental condition was represented by "BY %", since BY frequency always affected B priors in a monotonic fashion (see Figure 4, all three groups). Again, we found that the estimated B ratio (BY/BX) significantly predicted the best-fit B prior derived from RT and accuracy behavior (**Figure 29**, right panel).

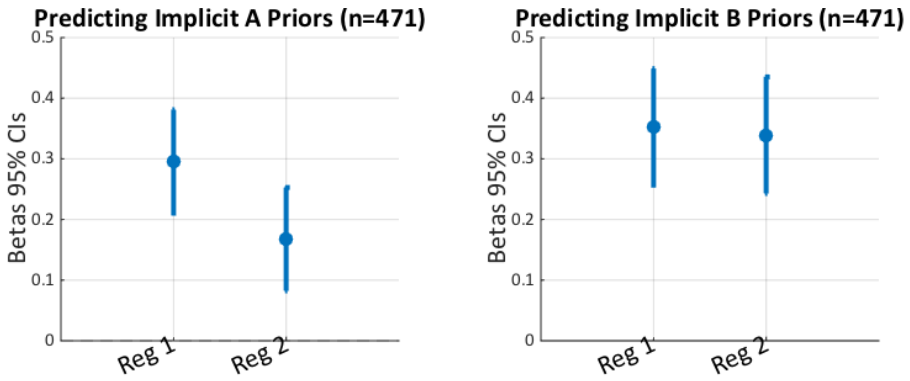


Figure 29 Best-fit implicit priors were significantly predicted by explicitly reported priors. The left panel shows the results of regressing the best-fit implicit A prior against the explicitly reported AX:AY ratio (Reg 1) and against "BY % * zscore(ABsimilarity)" (Reg 2), which was found to capture the effect of experimental condition on the best-fit A prior (Figure 6). The right panel shows the analogous regression for the best-fit B prior, as a function of the explicitly reported BY:BX ratio (Reg 1) and BY frequency (Reg 2). Both regressions revealed that the explicitly reported frequencies at debriefing could predict the implicit priors (derived from RT and accuracy data during the experiment) on a participant-by-participant basis.

Thus, while the implicit and explicit measures seemed to differ in the way response priors were grouped across cues, the two measures were nevertheless correlated across participants, such that a participant's explicit estimates of AX vs. AY frequencies could predict how much faster and more accurate they had been on AX vs. AY trials (for B priors).

Implicit priors are close to optimal given frequency estimates for M-turk but not Lab participants

Given the tight relationship between a participant's implicit priors during the experiment and their explicit reports at debriefing, we verified how far participants were from earning the maximal amount of reward rate on the task given the trial frequencies they reported explicitly. Recall that Agent 1 (**Figure 16**, left panel) was the agent that treated A and B independently and

used the starting points and thresholds that maximized reward rate given the true trial frequencies. Since human participants do not have access to the true frequencies, it is possible that the upper bound on how much reward they can earn in the task is set by how accurately they have learned the probabilities. Of course, there is no guarantee that the frequencies they reported explicitly were the same representations of probabilities that governed behavior on the task. Nevertheless, we verified whether some of the deviation from "optimality" (relative to Agent 1) could be explained by their explicit frequency reports.

To generate the "optimal priors" for Agent 1 displayed in **Figure 16**, we had calculated the cue-specific starting points and thresholds that earned the maximal reward rate given the true trial frequencies and given the average values of the other parameters (drift rate, cue noise and T0) across all our participants. This time, for each participant separately, we calculated the starting points and thresholds that earned the maximal reward rate given that participant's *reports* of the trial frequencies, and given the best-fit values of the other parameters (drift rate, cue noise, T0) we had obtained when fitting the DDM to the participant's behavior.

Figure 30 shows the priors that maximize reward rate given the true frequencies (row 1), the participants' estimated frequencies (row 2), as well as the best-fit priors (row 3) recovered when the starting points and thresholds are free parameters. Row 1 should technically be identical to the behavior of Agent 1 in **Figure 16** (left panel), except that rather than using the average parameters (drift rate, cue noise, T0) over the population, we predicted the optimal priors for each participant individually given that participant's best-fit drift rate, cue noise and T0. The bottom panels should technically be identical to the best-fit priors displayed in **Figure**

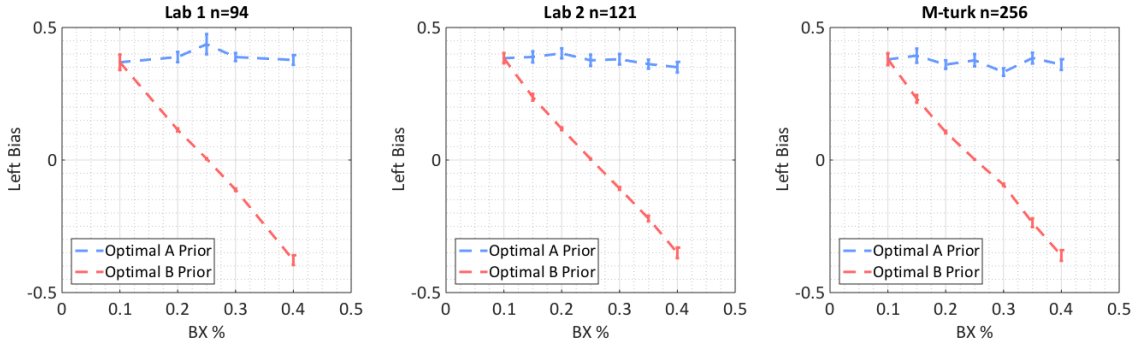
17, except that we show only the participants who chose to provide frequency estimates at debriefing, so the participants displayed in **Figure 30** are a subset of those in **Figure 17**.

Rows 4 and 5 in **Figure 30** show the overlay of the best-fit priors with the optimal ones. These comparisons between the best-fit and optimal priors revealed three main patterns, which are described in the three sections below.

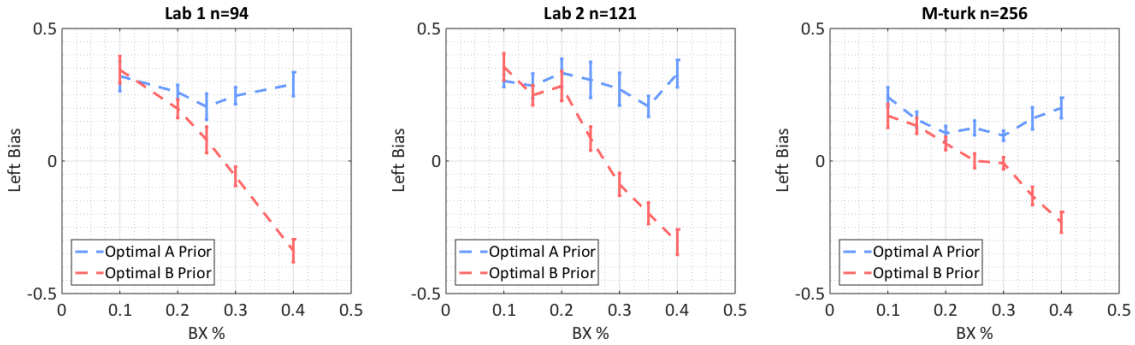
Estimated frequencies improve model fits relative to true frequencies

First, we see that the optimal priors given the *estimated* frequencies are much closer to participants' best-fit priors than the optimal priors given the *true* frequencies. In fact, the log likelihood of the data given the "optimal" parameters was consistently higher when the optimal prior parameters were found using participants' estimated frequencies (M=−167.0, SD=355.1), rather than the true frequencies (M=−188.9, SD=372.0; $p<10^{-14}$, z=7.8, as assessed by a two-sided Wilcoxon signed rank test.) In other words, putting the reported frequency estimates into the model significantly improved model fit quality.

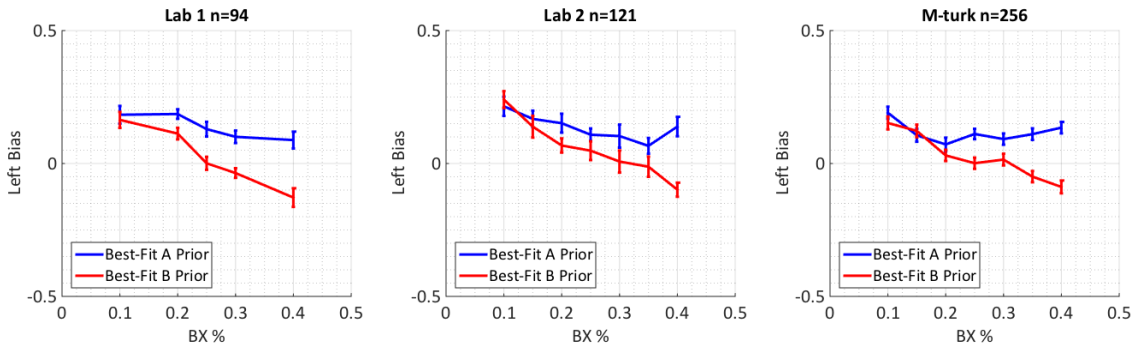
Row 1: Priors that maximize reward rate given true frequencies



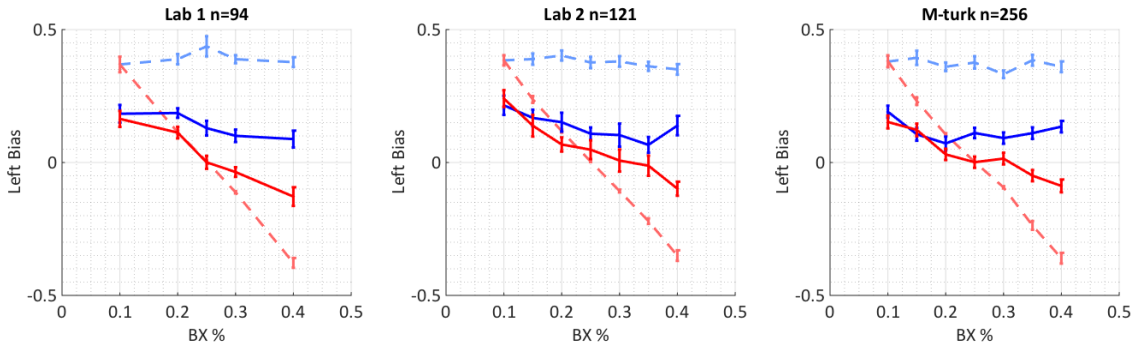
Row 2: Priors that maximize reward rate given estimated frequencies



Row 3: Best-fit priors



Row 4: Comparing best-fit priors and optimal priors given true frequencies



Row 5: Comparing best-fit priors and optimal priors given estimated frequencies

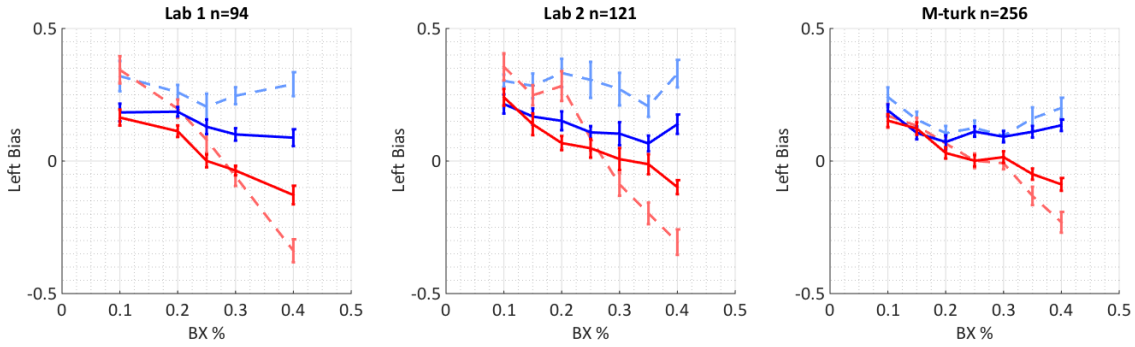


Figure 30 Comparing best-fit priors with priors that maximize reward rate given true frequencies and given estimated frequencies. We verified whether participants' behavioral priors, as measured from their RT and accuracy, were close to optimal given their estimates of the trial frequencies. For each participant, we calculated the starting points and thresholds that maximized reward rate given the best-fit values of the other parameters (drift rate, cue noise, T0) and given the true trial frequencies (row 1) or given the estimated frequencies (row 2) provided by the participant during debriefing. We found that participants' estimates of the frequencies provided significantly better model fits than the true frequencies, but this effect was significantly stronger for M-turk than for lab participants.

Best-fit priors deviate from optimal ones in systematic ways

Second, the best-fit priors for the M-turk group (right-most panel) appeared very close to the optimal priors given their frequency estimates, though this was not the case for the lab groups. To test the significance of this effect, we fit a linear regression model where we assessed whether the best-fit priors were significantly different from the optimal priors given the frequency estimates. We called this factor (whether the priors were derived from the optimal model or estimated from participants' data) the "model" factor, and tested whether the effect of model interacts with BX frequency (experimental condition, x-axis of **Figure 30**), the cue (A or B) and the experimental group (Lab 1, Lab 2, or M-turk). Our linear regression revealed significant overall interactions between model and cue (t=3.1, p=0.002), as well as

between model, cue and experimental condition (t=−4.5, p<10^{-5}). The model by cue interaction was present because the best-fit priors were lower than optimal for A, but higher than optimal for B (**Figure 30**, row 5). The model by cue by experimental condition interaction was present because the best-fit B priors were lower than optimal when BX frequency was low and higher than optimal when BX frequency was high, but this interaction was not present for A priors, for which the best-fit priors were always lower than optimal. Importantly, as is visible when comparing the Lab and M-turk groups by eye, the M-turk group significantly modulated the interaction between model and cue (t=−2.19, p=0.029), as well as between model, cue and experimental condition (t=2.72, p=0.007), since both of these interactions were much weaker in M-turk participants. The effects of the model are weaker in the M-turk group because the best-fit priors are quite close to the optimal priors given their frequency estimates. In fact, when we ran the same model only for the M-turk group, we found that the model factor was not significant in its interaction with cue (t=1.16, p=0.248), but did interact significantly with cue and condition (t=−2.4824, p=0.013). This is because the best-fit B priors were higher than optimal when BX frequency was high, while the best-fit A priors were slightly lower than optimal when BX frequency was high, whereas for all other conditions the best-fit and optimal priors were indistinguishable from each other on M-turk (**Figure 30**, bottom right panel).

To summarize, the linear regression showed that best-fit priors were significantly closer to optimal given estimated frequencies for M-turk participants than lab participants. It also revealed that, even though the frequency estimates explained some of the deviation from optimality, best-fit priors were still closer to 0 than they should have been.

M-turk starting points are indistinguishable from optimal when accounting for overly high decision thresholds

The above regression revealed that participants' best-fit priors were less extreme than the optimal priors predicted by a model that maximizes reward rate given participants' frequency estimates. For example, the best-fit B prior (solid red line) consistently had a lower slope than the optimal B prior (dashed red line); the slope was lower because the best-fit priors were always closer to 0 than the optimal ones (**Figure 30**, row 5).

One reason that the prior (starting point / threshold ratio) could be too low relative to a prior that earns the most reward rate is if the threshold is too high. Since human participants have repeatedly been shown to display an accuracy bias, using overly high thresholds relative to the optimal ones (Simen et al., 2010; Simen et al., 2011), we sought to verify whether all of the deviation from optimality could be caused by using overly high thresholds (rather than overly low starting points).

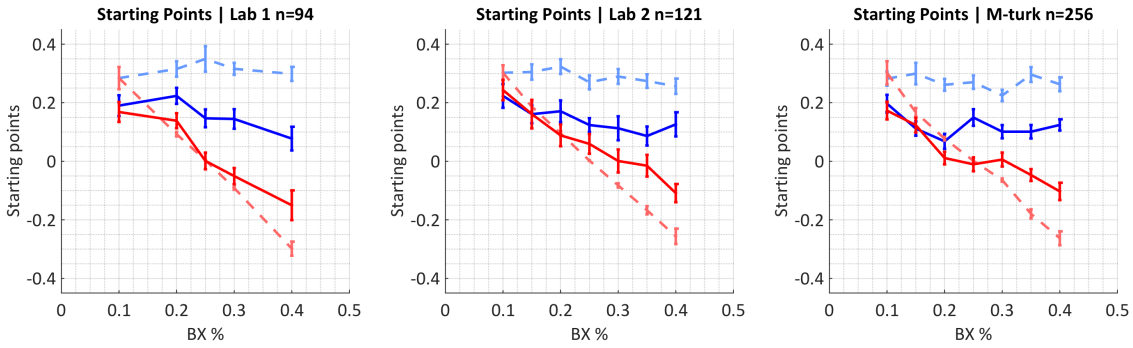
Figure 31 shows the same comparisons as **Figure 30**, plotted for starting points and thresholds separately. Rows 3 and 4 of **Figure 31** show that participants' best-fit thresholds were invariably higher than they should have been given their frequency estimates ($p<10^{-17}$, *z-value*=14.5 for A cues; $p<10^{-17}$, *z-value*=14.6 for B cues). They also reveal that the thresholds did not differ across cues (*p*=0.252, *z-value*=1.14), suggesting that we could have used a single threshold parameter, rather than allowing it to vary by cue. Rows 1 and 2 show that when the overly high thresholds were taken into account, the best-fit starting points were much closer to the optimal starting points given participants' estimated frequencies, and that for M-turk participants the two lines almost perfectly overlapped.

To quantify this, we repeated the above regression on the starting points alone, rather than the starting point / threshold ratios. Again, if the effect of "model" is significant, it suggests that best-fit starting points differ from optimal ones given estimated frequencies. We again found a significant interaction between model and experimental condition (t=2.05, p=0.040; because starting points deviated from optimal ones more for high BX frequencies) as well as an interaction between model, condition and cue (t=-3.22, p=0.001), because the starting points were too low for A cues and too high for B cues. Again, we found that this interaction was modulated by group, in the sense that it was significantly weaker in M-turk than lab participants (t=2.21, p=0.027). In fact, when we fit this regression only to M-turk participants, we could not find any significant effect of "model". The interaction between model, condition and cue, which was the primary way that differences between best-fit and optimal starting points manifested in lab participants, dropped below significance (t=-1.25, p=0.213) on M-turk.

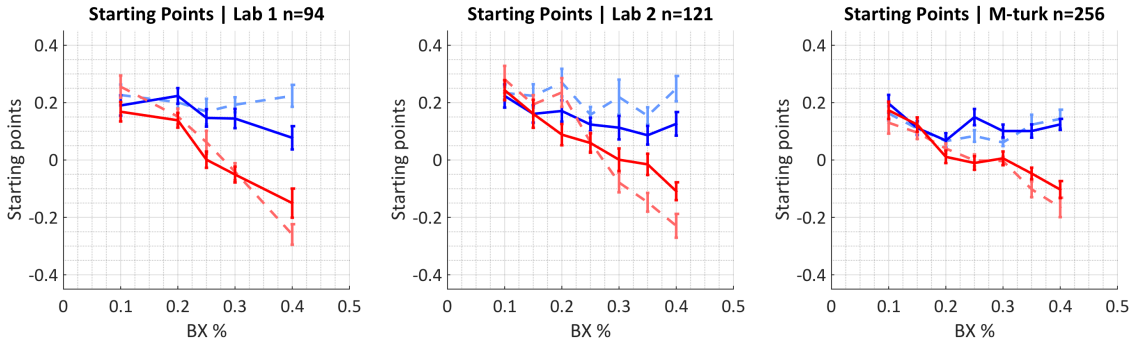
To summarize, the qualitative patterns in starting points were similar to the ones found in the starting point / threshold ratios. 1) For lab participants, starting points were closer to 0 than they should have been given their frequency estimates, especially as BX frequency increased. 2) For M-turk participants, starting points were so close to optimal given their frequency estimates that a statistical difference between them could not be detected.

In the Discussion, we propose possible explanation for this difference between lab and M-turk participants.

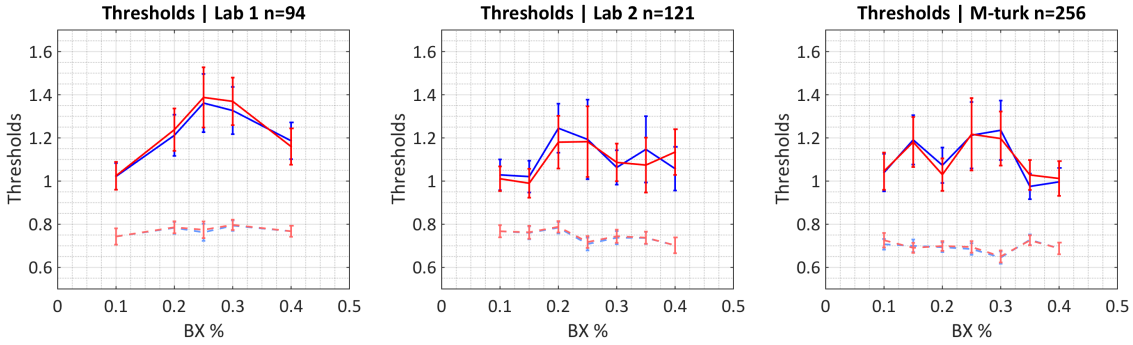
Row 1: Comparing best-fit and optimal starting points given true frequencies



Row 2: Comparing best-fit and optimal starting points given estimated frequencies



Row 3: Comparing best-fit and optimal thresholds given true frequencies



Row 4: Comparing best-fit and optimal thresholds given estimated frequencies

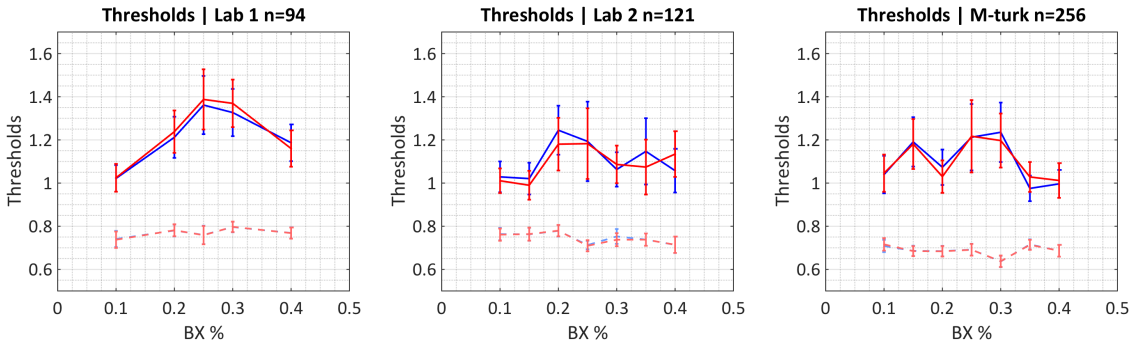


Figure 31 Comparing best-fit starting points and thresholds with optimal ones given the true and estimated frequencies. Rows 3-4 show that participants' best-fit thresholds were much higher than thresholds that optimize reward rate, replicating a previous result from the literature that participants tend to wait too long to respond, favoring accuracy but losing on reward rate. Row 2 shows that, once the overly high thresholds were accounted for, M-turk participants' best-fit starting points were indistinguishable from optimal given their frequency estimates. In other words, M-turk participants' "incorrect" frequency estimates appeared to account for all sub-optimality in their implicit priors. On the other hand, lab participants still deviated from optimality given their estimates, using starting points that were too close to 0.

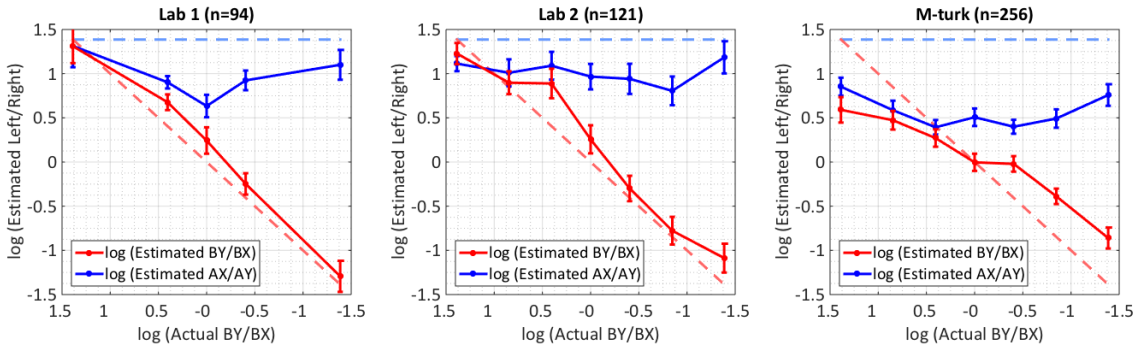
Implicit and Explicit Priors are Closer to 0 than Optimal

The above analyses revealed that participants' best-fit starting points were invariably smaller in absolute value (closer to 0) than those predicted by a model that maximizes reward rate (Agent 1, **Figure 31**, row 1) given the true trial frequencies.

When the participants' estimated frequencies are fed into this model (**Figure 31**, row 2), we begin to see a group difference, where lab participants' starting points remain closer to 0 than is optimal given their frequency estimates, while M-turk participants starting points are indistinguishable form optimal given their estimates.

One reason for this difference between groups is that the explicit frequency estimates themselves displayed a tendency to shrink response probabilities towards 50:50, but to a much greater degree in M-turk than in lab participants. This is evident in the reported AX:AY and BY:BX ratios displayed in **Figure 20**. (The figure is reshown below, for convenience.) Examining the M-turk reports first, we see that the reported AX:AY ratios (solid blue line) never reach a value higher than 1, while the log of the true ratio (dashed blue line) is 1.5. Similarly, the curve of reported BY:BX ratios (solid red line) displays a shallower slope than the true BY:BX ratios (dashed red line), such that the reported ratios are consistently closer to 0 than the true ratio.

Importantly, this was not the case in the frequency estimates reported by lab participants. Although the reported AX:AY ratio does deviate from the true ratio most of the time, we see that at its highest values (on the extremes of the curve), it does reach the true value of 1.5. This is even more clear from the reported BY:BX ratios (solid red line), where we see that the ratios are actually **further** from 0 than the true ratios in the domain when A and B are similar and being pulled towards each other.



Copy of Figure 20: Reported AX:AY ratios and BY:BX ratios across three participants groups. The dashed lines reflect the true frequencies that participants were given, while solid lines reflect their estimated frequencies. The AX:AY ratio reflects the expected ratio of left:right trials after A cues, while the BY:BX ratio reflects the expected ratio of left:right trials after B cues.

To quantify these effects, we performed a Wilcoxon signed rank test comparing the absolute values of the estimated BY:BX ratios with the true BY:BX ratios, separately for M-turk and lab participants. Absolute values were used to verify whether the reported ratios were always smaller in **magnitude** than the true ratios, regardless of whether the true ratios were positive (left-biased) or negative (right-biased). We performed this analysis only on BY:BX ratios, since we knew that reported AX:AY ratios significantly differed from the true value due to the merging of probabilities across cues (they were being pulled down towards B). We found that the reported B ratios (M=0.60, SD=0.55) on M-turk were significantly smaller in magnitude than the true B ratios (M=0.77, SD=0.48; $p<10^{-4}$, $z\text{-}value$=−4.2). On the other hand, the reported B ratios in lab participants (M=0.84, SD=0.58) were significantly larger than the true ratios (M=0.70, SD=0.51; $p < 10^{-3}$, $z\text{-}value$ = 3.5). This is because lab participants' estimated B ratios were being away from 0 and up towards A ratios in the domain where A was similar enough (Figure 7). We additionally tested whether M-turk participants reported smaller AX:AY ratios than lab participants. A Wilcoxon rank sum test (equivalent to a Mann-Whitney U-test) showed that AX:AY ratios reported on M-turk (M=0.65, SD=0.60) were significantly lower in magnitude than those reported by lab participants (M = 1.03, SD = 0.61; $p < 10^{-10}$, $z\text{-}value$ = 6.58). This provides further evidence that M-turk participants' explicit reports displayed the same kind of underestimation of left:right ratios as their implicit priors, measured by their best-fit starting points. On the other hand, lab participants did not tend to underestimate ratios in their explicit reports, but did show underestimation in their best-fit starting points.

To summarize, all participant groups displayed best-fit starting points that were smaller in magnitude (closer to 0) than those that maximize reward rate given the true stimulus frequencies. In practice, this means that during their performance of the task, their left responses were not sufficiently faster and more accurate than their right responses when left was much more common (and vice versa, when right responses were far more common then left, participants' right responses were not sufficiently faster and more accurate than their left responses).

M-turk participants also displayed the same tendency to underestimate left:right ratios in their explicit reports of trial frequencies, while lab participants did not. This partly explains why M-turk starting points were close to optimal given their frequency estimates, while lab starting points were significantly lower than optimal even given their estimates.

Overly low implicit and explicit priors may result from underestimating frequent events and overestimating rare events. This tendency has been found repeatedly in the literature on probability estimation (Zhang & Maloney, 2012), and possible explanations are reviewed in the Discussion.

Requirements for a model that could capture participants' implicit priors

Thus far, we have found two important ways in which participants' implicit priors deviate from the ones that maximize reward rate (Agent 1): 1) participants average their response priors across A and B cues, and 2) participants adopt overly low starting points (that are too close to 0) even in conditions when A and B predict the same response probabilities (when BX%=10%, see **Figure 31**, row 1).

Any model that attempts to explain how participants' priors deviate from optimal ones (such as the Agent 2 or Agent 3 models) must therefore account for both of these patterns. It is particularly important to notice that the two patterns make similar predictions for certain experimental conditions where averaging response probabilities across cues amounts to shrinking them towards 0. This is particularly true in the "Split" condition (**Figure 15**) when AX%=40% and BX%=40%, when A predicts the left response while B predicts the right response. Averaging response probabilities across cues would result in a prior of 0 (since left and right probabilities are equal overall), as would shrinking both priors towards 0.

On the other hand, the two patterns make distinct predictions in the "Merge" condition, when AX%=40% and BX%=10%, when both cues predict the same probability left response. In this condition, averaging priors across cues has no effect, since the two cues make the same predictions about the left response, yet we still see that the best-fit starting points are lower than optimal, suggesting that they are being pulled towards 0.

For this reason, a model that captures both patterns would have to be fit hierarchically to all experimental conditions at once (this way, conditions where the two patterns can be distinguished would inform the best-fit parameter values for conditions in which they cannot be distinguished.)

Alternative Interpretations and Additional Analyses

This section describes alternative interpretations of our results, as well as additional analyses we performed to test those interpretations.

Averaging stimulus instead of response probabilities

Two of our experimental findings suggested that, when A and B predict different response probabilities, participants do not form independent representations for A and B priors. First, in the conditions where B predicted the right response while A predicted the left (SPLIT conditions), we found that explicitly reported AX:AY ratios were anti-correlated with reported BY:BX ratios (this is evident in **Figure 23**). Second, in our follow-up experiment, participants who began with 200 trials of the SPLIT condition exhibited negative correlations between their implicit A prior and their implicit B prior across blocks within a participant (**Figure 27**). In the "Conclusion of Experiment" section, we mentioned that these negative correlations could have two interpretations: participants could have learned that A and B exhibit similar *response* priors but in opposite directions, or participants could have learned that A and B predict similar *stimulus* priors.

One way to disentangle whether the negative correlations in the explicit priors are caused by averaging along the response or the stimulus dimension is to examine participants' explicit reports separately for each trial type (AX, AY, BX and BY), rather than computing the AX:AY and BY:BX ratios. If the negative correlations are caused by learning opposite priors along the response dimension, then we might expect estimates of AX trials to negatively correlate with estimates of BY trials (both predict the left response), and for AY trials to correlate negatively with BX trials (both predict the right response). On the other hand, if the negative correlations are caused by grouping of stimulus probabilities, then we might expect AX trials to correlate

positively with BX trials (both predict X) and AY trials to correlate positively with BY trials (both predict Y).

Indeed, computing the correlations for each pair of trial types separately showed results that were more consistent with averaging of *stimulus* probabilities in the SPLIT condition. To quantify this, we performed a regression for each experimental condition, where we predicted either AX estimates or AY estimates as a function of BX and BY estimates (two regressors).

Figure 32 shows how the betas for the BX and BY regressor vary as a function of experimental condition (BX %).

In MERGE conditions where A and B predicted similar *response* probabilities (BX < 25%), AX estimates were positively correlated with BY (but not significantly with BX) whereas AY estimates were positively correlated with BX (but not significantly with BY). In other words, A and B estimates were positively correlated along the Left/Right dimension, but not significantly correlated along the X/Y dimension.

In SPLIT conditions, where A and B predicted similar *stimulus* probabilities (BX > 25%), AX estimates were positively correlated with BX (but not significantly with BY) whereas AY estimates were positively correlated with BY (but not significantly with BX). In other words, A and B estimates were positively correlated along the X/Y dimension, but not significantly correlated along the Left/Right dimension.

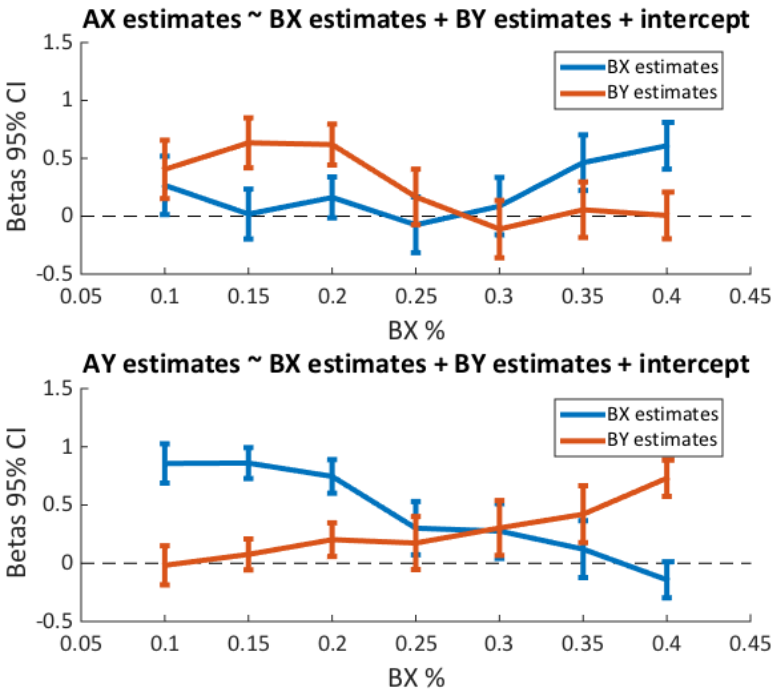


Figure 32 Explicit probability estimates were correlated along the response dimension in the MERGE condition and along the stimulus dimension in the SPLIT condition. To test whether the negative correlation between reported AX:AY and BY:BX ratios was caused by negative correlations along the response dimension, or positive correlations along the stimulus dimension, we correlated AX (and AY) estimates separately with estimates for BX and BY trials. A linear regression was performed separately for each experimental condition (BX %) where AX (or AY) estimates were regressed against BX and BY estimates. In SPLIT conditions (BX > 25%), we found that AX was positively correlated with BX and AY was positively correlated with BY, suggesting A and B estimates were correlated along the X/Y dimension. On the other hand, AX was not significantly correlated with BY and AY was not significantly correlated with BX, suggesting that A and B estimates were not significantly negatively correlated along the response dimension.

These results from the explicit reports speak strongly in favor of the interpretation that negative correlations between A and B response priors are caused by positive correlations along the X/Y stimulus dimension, rather than negative correlations along the response dimension. As BX % increases, A and B predict more different response probabilities (and more similar stimulus probabilities), a qualitative change in the way participants learn to group

observations across cues seems to occur. Participants appear to change from grouping *response* probabilities across cues to grouping *stimulus* probabilities across cues.

However, there is one aspect of the data that does not fit with the interpretation that participants are averaging stimulus probabilities across cues in the SPLIT conditions (when BX > 25%). Recall that Agent 3, which stops to average response probabilities when A and B make different predictions, predicts a U-shaped curve of A as a function of B (**Figure 16**). A is originally pulled down towards B, but then returns to its original level as B predictions become too dissimilar. While we did not see evidence for this U-shaped relationship in implicit priors (**Figure 17**), it was present in participants' explicit reports (**Figure 20**). However, the above interpretation suggests an additional model of behavior. An agent that averages stimulus probabilities when BX > 25% would behave differently from Agents 2 and 3. In those conditions, averaging stimulus probabilities across cues should cause *overestimation* of the right bias on B trials (**Figure 33**, top left panel), whereas Agents 2 and 3 predict *underestimation* of the right bias on B (**Figure 33**, top right panel). This is because BX frequency is always smaller than or equal to 40%, whereas AX remains at 40% in all conditions. If the probability of X is averaged across cues, then all conditions in which BX < 40 % should exhibit overestimation of BX trials and underestimation of AX trials. On the other hand, Agents 2 and 3 average the probability of the left response across cues. Again, since BY frequency (= the probability of left responses after B), is always smaller or equal to 40%, then averaging it with AX should cause overestimation of BY trials and underestimation of AX trials.

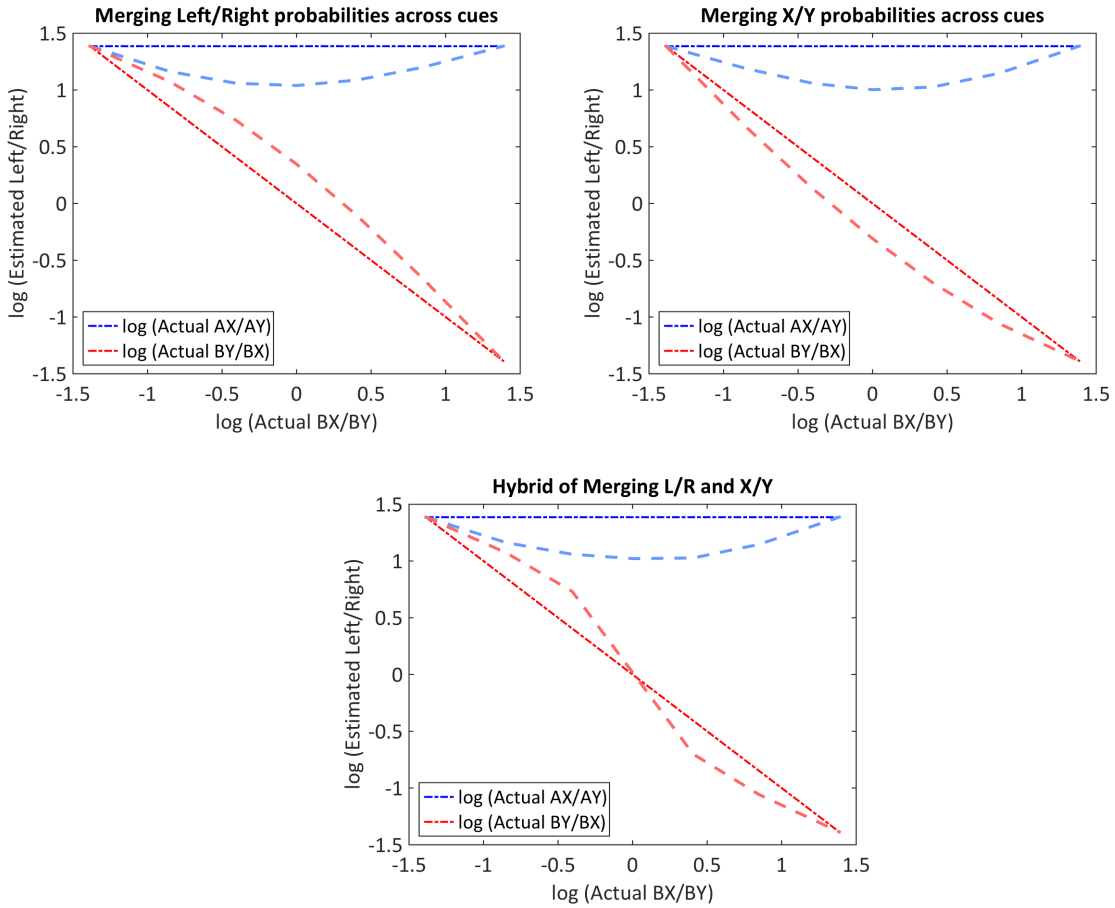


Figure 33 Predicting distortions in probability estimates if participants average left/right probabilities or X/Y probabilities across cues. Since BX frequency is always lower or equal to AX frequency (40%), averaging the X probability across cues causes *overestimation* of BX trials (greater right bias on B) and underestimation of AX trials (smaller left bias on A) (**top right panel**). Since BY frequency is also always lower or equal to AX frequency, averaging the left probability across cues causes *underestimation* of BX trials (greater left bias on B) and underestimation of AX trials (**top left panel**). Thus, the two modes of averaging make opposite predictions for distortions in BY:BX estimates. Finally, the **bottom panel** shows a hybrid strategy, which averages left/right probabilities when A and B make similar response predictions, but averages X/Y probabilities when A and B make similar stimulus predictions.

Figure 33 shows the resulting distortions in the estimates of AX:AY and BY:BX ratios under these two strategies (averaging left/right probabilities vs. averaging X/Y probabilities). The bottom panel shows a hybrid strategy that averages over the response dimension when the

cues predict the left response and averages over the stimulus dimension when the cues predict the X stimulus.

These predictions pose an important challenge to our interpretation that participants are averaging stimulus probabilities across cues, since they should then show *overestimation* of the right bias on B trials (the B prior should be *below* the optimal curve). However, participants' implicit and explicit B priors, both in lab and M-turk participants, have consistently shown *underestimation* of the right bias on B (such that the B prior is always *above* the optimal curve. See **Figure 17** and **Figure 20**).

One important confound that could be causing underestimation of the right bias on B is overall underestimation of probability ratios. As noted in the section "Implicit and Explicit Priors are Closer to 0 than Optimal", in addition to merging response probabilities across cues, participants deviated from optimality by underestimating probability ratios in general, such that both left and right biases were smaller than they should have been. To visualize the effect of underestimating probability ratios, we reproduced the plots in **Figure 33**, but this time, the logs of all probability ratios were multiplied by 0.7 (to mimic underestimation of probability ratios). The resulting plots in **Figure 34** show that when underestimation of probabilities is strong enough, the BY:BX estimates curve becomes shallower and is not sufficiently below the true value to be able to detect overestimation of BX frequency (right bias on B). Since the shallow curves in **Figure 34** mimic those we see in human behavior, future work will need to quantitatively model the effect of underestimation of probability ratios, merging left/right probabilities and merging X/Y probabilities across cues all in one model. Only then can we hope

to disentangle how much each of them affects the mean curves of A and B priors as a function of condition.

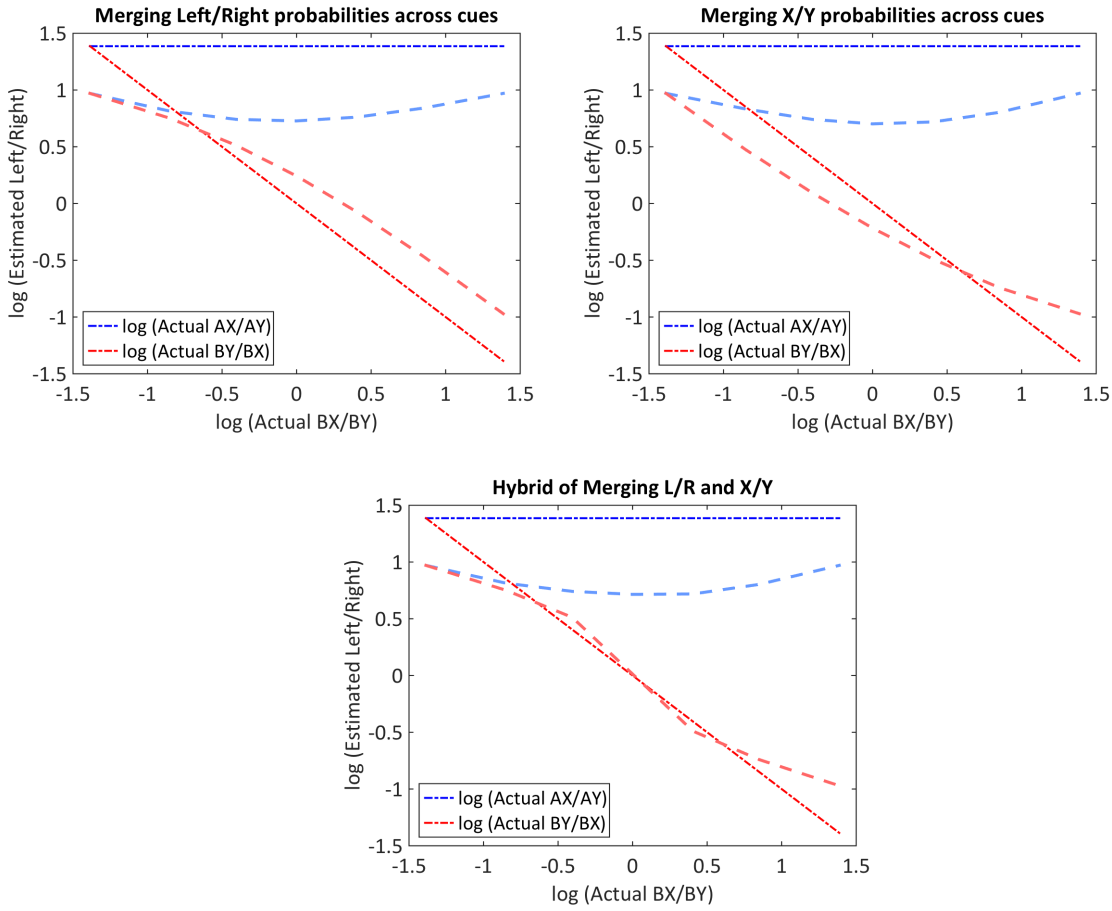


Figure 34 Predicting distortions in probability estimates if participants underestimate logs of probability ratios, in addition to averaging left/right probabilities or X/Y probabilities across cues.

What role do switch costs play in our implicit measure (accuracy and reaction times)?

The drift diffusion model we have fit in order to measure people's priors assumes that accuracy and reaction times on AX vs. AY trials (in the case of A priors) or BX vs. BY trials (in the case of B priors) should be the same regardless of which trial came just before. However, the

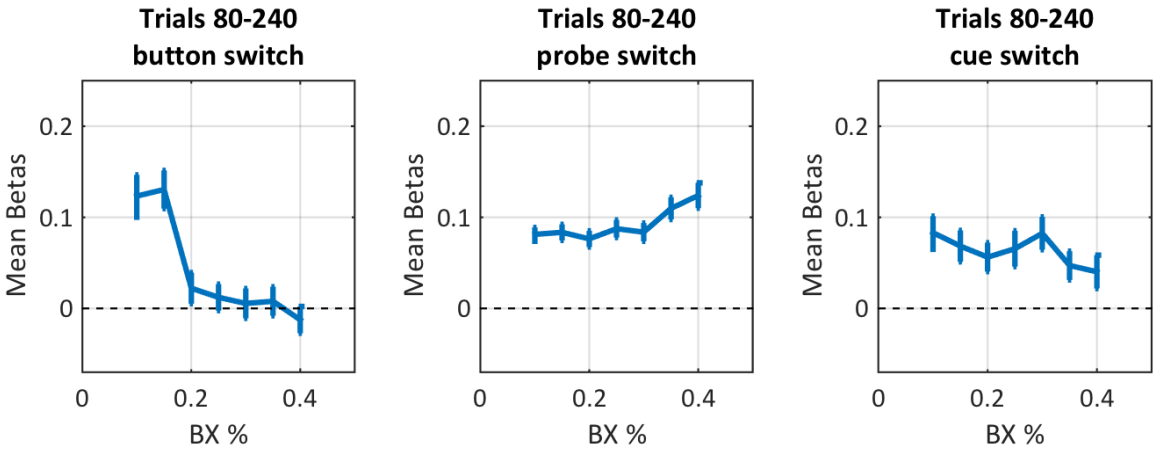
literature on task switching has shown that both reaction times and error rates should be lower if the current trial is a repeat trial (Kiesel et al., 2010). In other words, participants should be faster to respond on an AX trial if it was preceded by an AX trial, even if they have not learned any prior probabilities whatsoever. Moreover, several studies have shown that repeating a motor response causes a decrease in reaction time and error rate, but only when the task is also repeated; on the other hand, when the task switches, response repetitions are actually slower than response switches (Quinlan, 1999; Kleinsorge, 1999; Meiran, 2005; Rogers & Monsell, 1995; Sohn & Carlson, 2000). In our case, since A and B can be considered as "tasks" (the cue maps stimuli to motor actions), a left response on AX might be slower if it is preceded by a left response on BY than if it is preceded by a right response on BX.

Since our experimental conditions manipulate the probabilities of each of the four trial types, they also necessarily change the probability of switch and repeat trials. For instance, since AX trials were more frequent than AY trials, a participant was more likely to experience AX repetitions than AY repetitions. Since repetitions tend to be faster and more accurate than switches, it is possible that the difference we measure between AX and AY reaction times or accuracy is due to the difference between switch and repeat trials, rather than to probability learning.

To quantify the effect of switches vs. repeats in our data, we regressed each participant's time course of correct reaction times against three predictors: response button (left/right) switches, probe (X/Y) switches and cue (A/B) switches (as well as an intercept term). Each regressor was binary and contained a 0 if the current trial was a repeat and a 1 if the current trial was a switch. Thus, the response button regressor would have a positive coefficient if

response button switches were generally slower than response repeats. After fitting the regression separately to each M-turk participant's RT time course, the average betas for each of the three regressors are plotted in **Figure 35**, as a function of experimental condition (the frequency of BX/BY trials).

We found that response button switches were slower than repeats only in MERGE conditions (BX < 20%), in which AX and BY trials are most common and where participants press the left button most of the time. This is not surprising and suggests that the more a condition involves pressing the same button, the more difficult button switches become. In contrast, we found that probe (X/Y) switches were slower than repeats in all conditions, suggesting that stimulus repetitions influenced behavior more reliably than response repetitions. This effect was strongest in SPLIT conditions (BX > 30 %), where AX and BX trials are the most common and where participants see X most of the time. Again, this result is sensible and suggests that the more a condition involves seeing the same probe (X), the more difficult probe switches become.



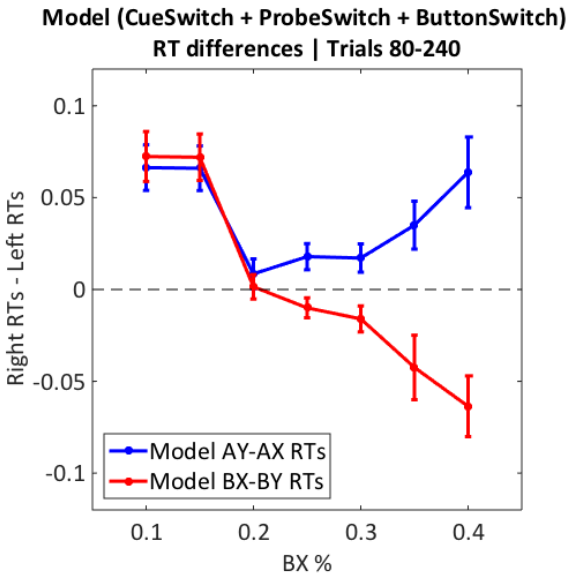


Figure 35 The cost of button switches and stimulus switches varies with BX frequency. Each M-turk participant's time course of correct response times (RTs) was regressed against three predictors: response button (left/right) switches, probe (X/Y) switches and cue (A/B) switches. Each regressor was binary, equal to 0 if the trial was a repeat and 1 if the trial was a switch. The **top row** displays the average regression betas (over participants) and standard errors of the mean. Trials where the probe switched were generally slower than probe repeats, though this effect grew as BX frequency increased. This suggests that switching probes is more difficult in conditions where X is a lot more frequent than Y, as opposed to conditions where X and Y are equally likely (BX=10%). Similarly, response button switches were slower than repeats only in conditions where AX and BY are common and the left response is more likely. This suggests that switching responses is more difficult in conditions where left responses are a lot more frequent than right responses, as opposed to conditions where left and right are equally likely (BX=40%). Finally, the **bottom row** displays simulated RTs from the regression model, generated from a weighted sum of the three regressors and intercept term, weighted by their coefficients. Average AX, AY, BX and BY RTs were calculated from this simulated time course, and the differences between AY-AX RTs (blue) and BX-BY RTs (red) were plotted using the same conventions as **Figure 18**. These simulations show that a model with higher probe switch costs in SPLIT conditions and higher response switch costs in MERGE conditions tends to reproduce the U-shaped curve of A as a function of B.

Having obtained the coefficients for each regressor, we could then generate fake RT time courses based solely on these regressors. For each participant, we generated simulated RT time courses that were a weighted sum of the three regressor and intercept term, weighted by their respective coefficients. Using these simulated RTs, we could regenerate the same plots as in **Figure 18**, showing how much slower the model was on right than left responses after A vs. B cues. We were amazed to discover that a regression model constructed only from those three predictors fit to M-turk data could produce behavior that remarkably resembled M-turk participants' implicit priors *and* their explicit frequency estimates. Specifically, the model produces a U-shaped curve of A prior as a function B.

At first glance, these results could cause concern that the patterns we see in behavior can be entirely explained by the fact that repeats are faster than switches and by the fact that repeats get even faster as the number of repeats increases. The problem with such an interpretation is that it suggests that the apparent merging between A and B response priors and stimulus priors was not due to structure learning over the course of blocks, but rather to the facilitatory effects of trial repetitions, which occur on the timescale of a handful of trials.

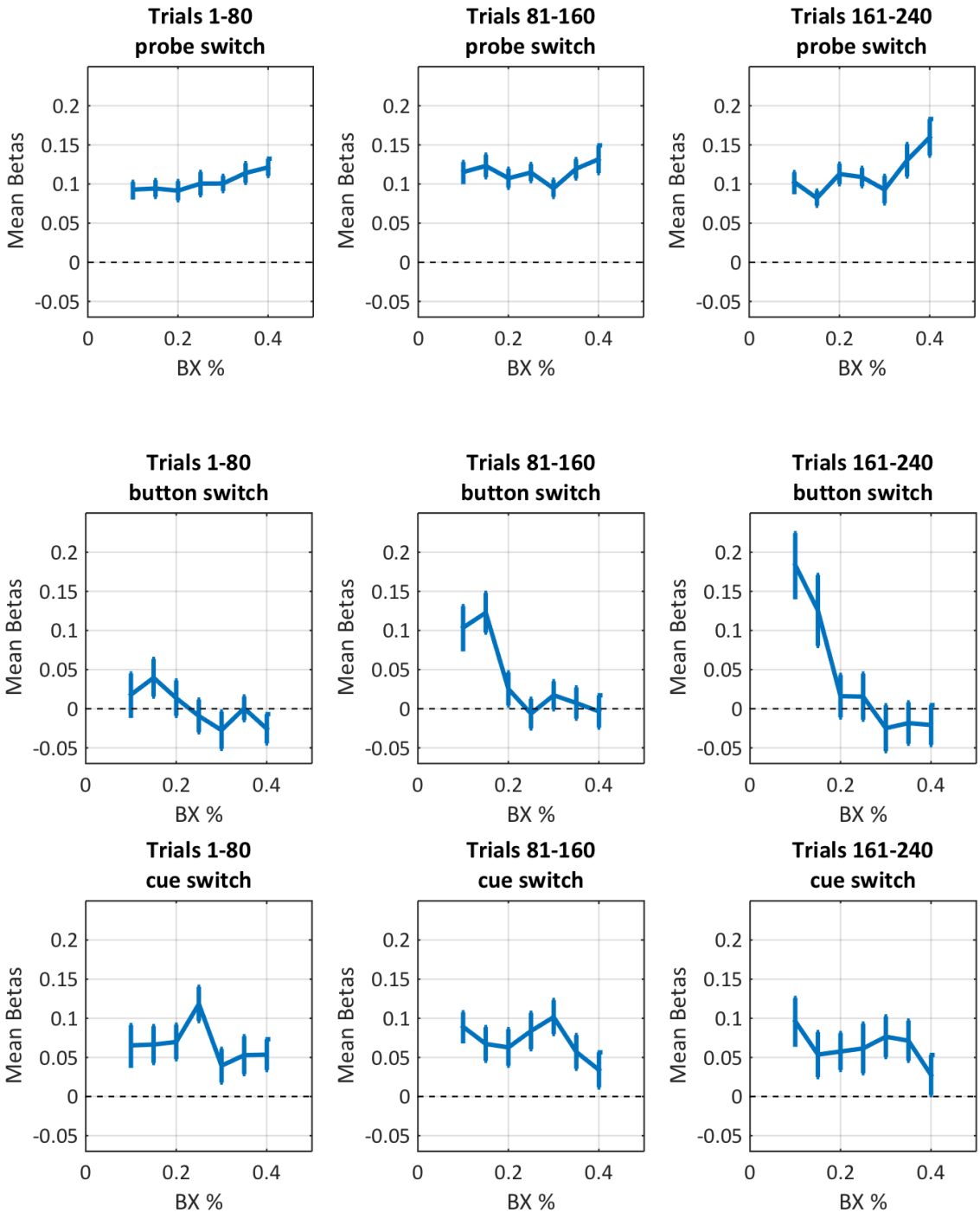
If the effect of switches and repetitions is on the time scale of a few trials, we should be able to see it starting from the very first blocks, since each block of 40 trials was constrained to have the exact same trial frequencies. Fitting the regression model separately to the first 80, second 80 and last 80 trials of the task revealed that these patterns grow tremendously from the first 80 to the last 80 trials of the experiment. **Figure 36** (row 1) shows that, for conditions where BX > 30%, where X is more likely overall, the cost of switching probes increases substantially over the course of the experiment. Similarly, for conditions where BX < 20%,

where left is more likely overall, the cost of switching response buttons increases substantially over the course of the experiment (**Figure 36**, row 2). This shows that while switch costs captured a large component of participants' behavior, the effect of switches changed gradually over time, as participants learned the probabilities of the trial types. In other words, it appears that switch costs change substantially as participants learn the statistics of the task, and may themselves be a reflection of structure learning.

This is line with the time courses of A and B priors in the follow-up experiment (**Figure 25**); fitting the DDM to each block of 40 trials separately showed that the priors increase in strength over the course of blocks. Both of these findings suggest that our results cannot be explained fully by switches vs. repetitions on the scale of a few trials, and are influenced by statistical learning on the scale of hundreds of trials.

Further evidence that our results are caused by learning on a longer time scale is provided by the explicit reports of trial frequencies, which show similar patterns to participants' RT and accuracy behavior on the task, despite the fact that they are provided by entering numbers at debriefing, after the task is complete.

Finally, the fact MERGE participants learn higher response button switch costs over time, while SPLIT participants learn higher probe switch costs over time, provides further support for our interpretation that participants learn to group observations along the response dimension in the MERGE conditions, but learn to group them along the stimulus dimension in the SPLIT conditions.


Trials 1-80
probe switch
Trials 81-160
probe switch
Trials 161-240
probe switch
Trials 1-80
button switch
Trials 81-160
button switch
Trials 161-240
button switch
Trials 1-80
cue switch
Trials 81-160
cue switch
Trials 161-240
cue switch
Mean Betas
BX %
0.2
0.15
0.1
0.05
0
-0.05
0
0.2
0.4

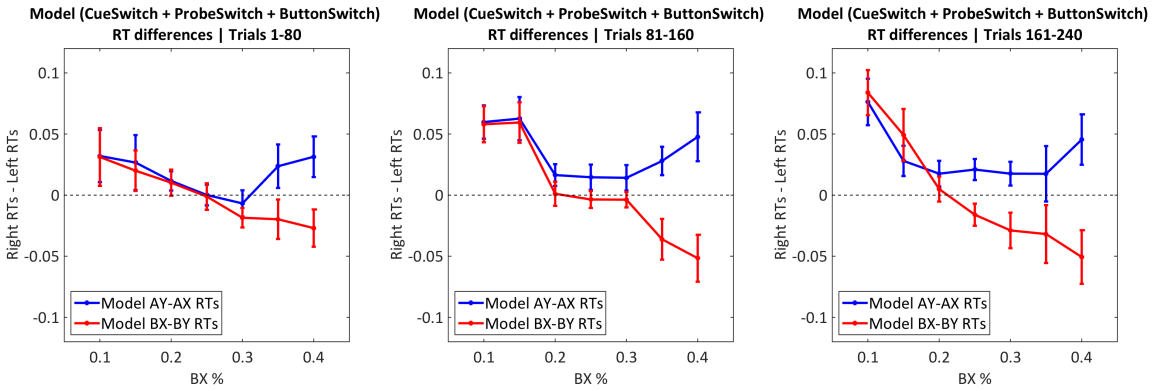


Figure 36 Evolution of response switch costs, probe switch costs and cue switch costs over blocks. The same regression model as above was fit separately to trials 1-80, trials 81-160 and trials 161-240, dividing the experiment into three chunks of 80 trials. The **top row** shows that the average betas for the probe switch regressor increase substantially over blocks, but only for conditions where BX > 30%, such that the difference in probe switch costs across conditions becomes more pronounced with time. In other words, in conditions where the X probe is far more likely than the Y probe, probe switches become more costly over the course of the experiment. Similarly, the **second row** shows that the average betas for the button switch regressor increase substantially over blocks, but only for conditions where BX < 20%, where the left response is far more likely than the right. In other words, switching response buttons becomes more costly over the course of the experiment in those conditions. The **bottom row** shows the difference between simulated RTs, generated from the regression model by taking a weighted sum of the three regressors multiplied by their coefficients. While the U-shaped pattern of A as a function of B is present even in the first 80 trials, this pattern changes qualitatively over the course of blocks.
These changes in the way the regressor betas vary by experimental condition show that the patterns shown in **Figure 35** (when fitting the model to the entire experiment) are a result of learning over the course of blocks, rather than produced by the frequency of repeats and switches on the time scale of trials.

Cue noise, error rates and reaction times as a function of condition

One of the goals of our study was to test whether people are able to merge predictions about upcoming stimuli or appropriate responses across cues, even in a task where differentiating the cues is necessary for making the appropriate response to the stimulus. Our previous analyses suggest that participants might be grouping response probabilities in the

MERGE conditions, yet grouping stimulus probabilities in the SPLIT conditions (**Figure 32**). If combining probabilities across cues comes at the cost of differentiating the two cues for the purpose of responding in a cue-dependent manner, then we might expect overall error rates or reaction times (RTs) to increase as the two cues make more similar predictions, either on the response or stimulus dimensions.

In addition to overall RT and error rates, we were interested in the best-fit values of the cue noise parameter, derived from fitting the DDM to participant behavior. As the cue noise parameter increases, the model increasingly confuses A and B cues, using the A diffusion parameters on B trials and vice versa. These diffusion parameters include the correct response rule (Left on X vs. Right on X) and correct set of priors. Thus, as cue noise increases, the model begins to confuse the A and B response rules and priors. If merging probabilities across cues also leads to greater confusability of the cues, then we should see higher cue noise as A and B make increasingly similar predictions.

To test this, we regressed cue noise, error rates, and mean correct reaction times against two predictors:

$$CueNoise \sim BY\% + zscore(BY\%) * zscore(ABsimilarity)$$

The first regressor was simply the experimental condition, while the second regressor followed a U-shaped curve. If cue noise, error rates or RTs are higher when the cues make more similar predictions along either the response or stimulus dimensions (when BX is very low or when BX is very high), then the U-shaped curve regressor should have significantly positive coefficients.

We found that none of the measures varied significantly with BY frequency or with the interaction between BY frequency and ABsimilarity (**Figure 37**). If anything, the regression coefficients tended to be negative. For the BY % regressor, a negative coefficient suggests that cue noise, error rates and RTs are lower when BY % is high, suggesting that far from impairing performance, the condition in which A and B make similar response predictions might actually facilitate performance. Examining the plots of cue noise, error rates and correct RTs (**Figure 37**) suggested that for the Lab 1 group of participants, who experienced 480 trials of the task, the BX=25% BY=25% condition actually seemed to have the highest error rates and cue noise, whereas the BX=10% (MERGE) condition seemed to have the lowest. In other words, participants performed equally well, if not better, in the MERGE condition, when A and B made the same response predictions, and appeared to perform worse (though not significantly), when B did not make any predictions about the stimuli or responses.

Finally, we verified whether individual differences in how much participants separated their implicit A and B priors were correlated with performance. Separation between A and B response priors was quantified as the difference between the best-fit A prior and the best-fit B prior. We did not find any significant relationship between prior separation and cue noise (95% CI=[-0.10, 0.08]), error rates (95% CI=[-0.08, 0.10]) or correct RTs (95% CI=[-0.09, 0.08]), nor any significant interaction between prior separation and experimental condition.

All of these results provide support for the interpretation that participants are able to combine both response and stimulus probabilities across cues, while still differentiating the cues for the purpose of making the appropriate (context-dependent) response to the probe stimulus. If anything, the data trend in the opposite direction, such that increasing similarity

between cue predictions appears to improve task performance. (It is possible that it improves task performance because greater similarity between cues causes greater compression, such that participants have to maintain less information about the differences between the two cues.)

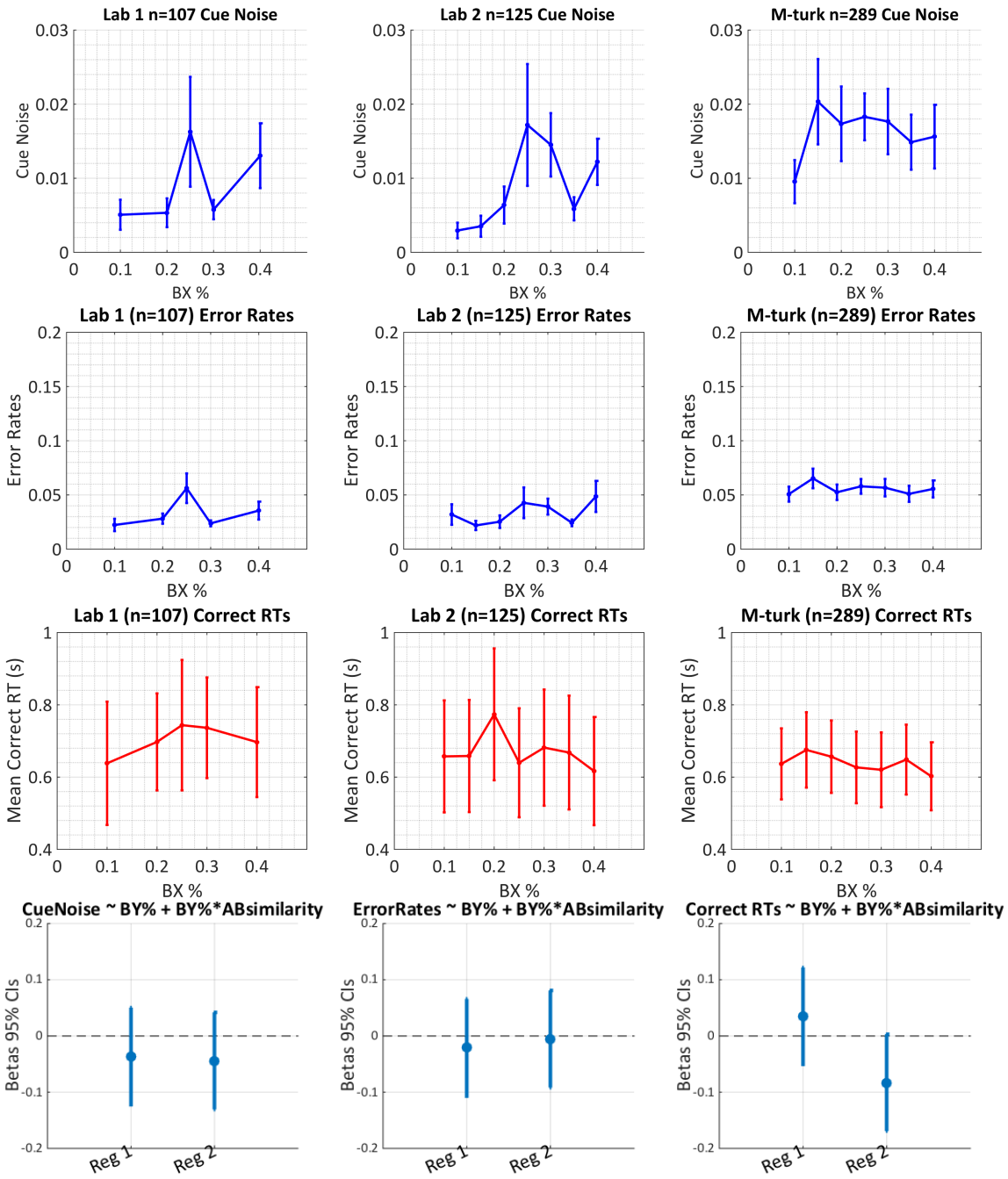


Figure 37 Cue noise, error rates and reaction times did not vary significantly with BX frequency. After pooling the data from all participants into a single regression model, we verified whether the cue noise, error rates or mean correct RTs for each participant varied linearly with experimental condition (BY %, Regressor 1) or followed a U-shaped curve (zscore(BY %)*zscore(ABsimilarity), Regressor 2). Neither regressor significantly predicted any of the measures, suggesting that participants were able to maintain more or less the same level of error rates, RTs and cue noise, regardless of whether the cues made similar response predictions.

Merging appears graded, rather than binary, both across and within participants

All of the results above hinted that combining response or stimulus predictions cues was a graded, rather than binary, process. The degree to which A was influenced by B changed gradually with BX frequency (the degree of B similarity), and the correlations between A and B priors across participants varied gradually from positive to neutral or negative (in the implicit vs. explicit measures) across conditions, with no indication of a step-like jump from grouping to not grouping (or from positive to negative correlations.) Nevertheless, it was possible that in each BX frequency condition, some proportion of participants had "perfectly" merged and some proportions of participants had "perfectly" split their priors, and these bimodal distributions would not be reflected in the averages if the number of perfect splitters and perfect mergers varied gradually across conditions. For this purpose, we examined whether the distribution of differences between A and B priors exhibited any signs of bimodality or multi-modality across participants. **Figure 38** (left panel) shows the distribution of differences between implicit A and B priors across all lab and M-turk participants and all conditions from the original experiment. The distribution of prior differences across participants shows no hint of multi-modality (Hartigan's dip statistic = 0.0094, bootstrapped p-value=0.997).

However, bimodal distributions might not be visible across participants if the step-like jump from merging to splitting occurs within participants. Specifically, participants might always begin the experiment with merged priors, but eventually gain enough evidence that the cues make distinct predictions and then adopt two different parameters. To check for multi-modality within participants, we also examined the distribution of prior differences for the follow-up

experiment, in which we had fit 10 blocks of 40 trials separately for each participant. For each block separately, we subtracted the best-fit B prior from the best-fit A prior, and then pooled the prior differences across blocks and participants, resulting in 10 blocks * 83 participants = 830 difference values. The distribution of prior differences both within and across participants (**Figure 38**, right panel) did not show any hint of bimodality either (Hartigan's dip statistic = 0.007, bootstrapped p-value=0.9998). To be sure that pooling these differences across participants did not obscure within-participant effects, we also performed the Hartigan's DIP test within each participant across the 10 blocks. Despite performing the test 83 times, we found that only 1 out of 80 participants exhibited an uncorrected p-value of less than 0.05. Thus, the distribution of prior difference appeared unimodal in 99% of participants.

These results, combined with our other analyses, suggest that even within participants, the degree of merging response predictions across cues varies gradually over time. This is counter to the idea that participants either entirely merge or split predictions across cues, which would predict a jump from merged to separate response priors once a participant had reached threshold and was confident that the two cues made different predictions.

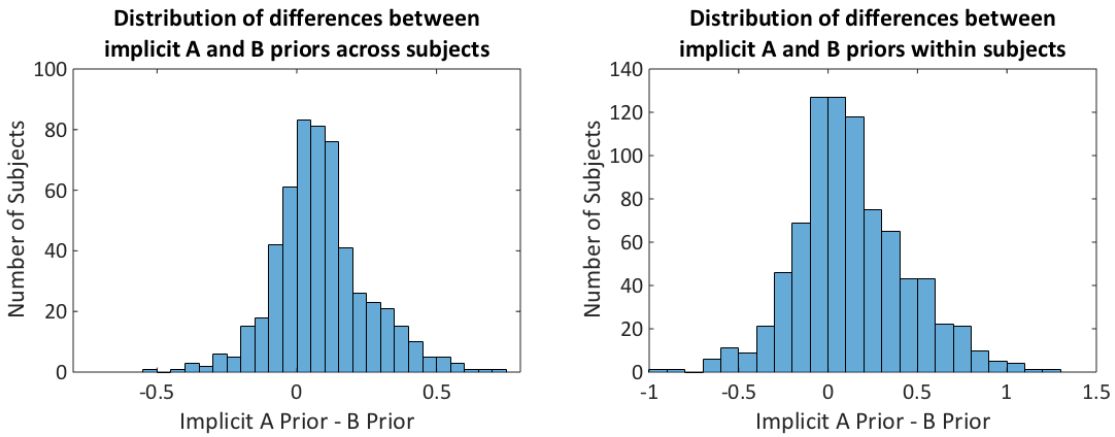


Figure 38 Difference between A and B priors did not exhibit any bimodality across participants or across blocks within participants. We verified whether the difference between participants' A and B priors might preferentially inhabit two extremes: 0 (suggesting that they were merged) or a large distance (e.g. 0.5), suggesting that they were split. Bimodality in the distribution of A prior – B prior differences would suggest that splitting was an irreversible process of creating a new representation, and that there was some threshold beyond which cue predictions were considered sufficiently different. However, we did not find any evidence of bimodality either across participants (**left**) or across blocks within participants in the follow-up experiment (**right**). Our results favor the interpretation that the priors are never fully split or fully merged, and that the relationship between them varies in a graded fashion with probability learning.

Discussion

Adaptive behavior relies not only on our ability to act appropriately in a given situation but also on our ability to predict which events are more likely in a certain context. If the probability of a certain event (for example, running into a colleague) differs substantially by context (grocery store vs. lab), then learning context-dependent probability distributions could substantially increase the accuracy of our predictions and our preparedness for a given event. However, if our observations are limited (either due to small amounts of data or noisy measurement), it may not be clear whether the probability of an event is different between two contexts. If the underlying probability distributions for two contexts are similar, then combining and averaging them could actually improve robustness to noise. Additionally, forming expectations that do not depend on context can be advantageous for making predictions in situations we have not previously encountered (i.e., generalization). Thus, developing separate predictions for each context may not be feasible in complex real-world situations (in which the effective number of contexts is infinite) and may not be advantageous (in environments with consistent structure). Nevertheless, some amount of context-specificity is necessary when a context predicts a substantially different probability of a given event, compared with other contexts.

Previous literature suggests that participants might resolve this trade-off in a Bayesian way, combining predictions across contexts when they are similar enough, but forming context-specific predictions after accumulating enough evidence that the probability differs by context.

As reviewed in the Introduction, such non-monotonic grouping of predictions across cues has been shown in both object perception (when people combine information across cues in a visual scene to determine a property of an object; Girshick & Banks, 2009) and perceptual estimation (when people learn an association between a seemingly irrelevant perceptual feature, such as color, and another feature of behavioral relevance, such as the number of circles; Gershman & Niv, 2013).

Unlike sensory cue combination studies, in which merging across cues occurs within individual trials, the Gershman & Niv (2013) study resembles ours in several ways. Two perceptual cues (colors) predicted the probability of another perceptual attribute (number of circles on the screen) that was relevant for behavioral performance. Rather than presenting the two cues simultaneously, only one color was shown on a given trial, such that the association between colors and the number of circles was learned across many trials. However, one key difference remains between our study and the Gershman & Niv experiments, as well as studies of cue combination: in those paradigms, the perceptual features that were integrated to make predictions about the world could usually be discarded once integration had occurred. In Gershman & Niv, the color was not mentioned in the instructions, and while participants might learn that color was *informative*, it was not crucial for performing the task correctly. In cue combination studies, perceptual features are integrated for the purpose of forming a unified and maximally accurate representation of the environment, and behavior is usually driven by that resulting percept, rather than by each feature separately. In such paradigms, people should be able to behave adaptively even if information about individual cues is lost, and might

even gain in performance by using only the unified (averaged) percept to guide behavior (Ernst and Bülthoff, 2004).

Our study set out to investigate whether probability distributions predicted by two cues can be "merged" even when differentiating the cues is crucial to behaving appropriately. While the colors in the circles task or the features in cue combination experiments are used to guide a perceptual decision (the number of circles on screen, or the slant of a grating), it is only the result of this perceptual decision that is then used to guide responding. On the other hand, in the AX-CPT, the A and B cues not only predict which stimulus is more likely to appear next (X or Y) and which response is more likely (left or right), they also deterministically flip the correct response to the stimulus. In other words, the response rules signaled by the two cues must remain distinct in order to guide behavior, even if stimulus or response probabilities are merged across cues over the course of trials.

Our results are entirely consistent with weak fusion of cue predictions in the domain of context-dependent decision-making. Recall that "weak fusion" describes strategies in which both individual cue information and the merged representation (which results from averaging predictions across cues) are retained and used to guide behavior (Landy et al., 1995). First, we found that implicit A and B response priors (derived from fitting the DDM to RT and accuracy data) were pulled towards each other to the degree that A and B made similar predictions about the response ("Regression: Quantifying the effect of BX/BY frequencies on best-fit A prior" section), and that across participants, the degree to which a participant's B prior predicted their A prior increased significantly as A and B made increasingly similar predictions.

Examining the distance between A and B priors within and across participants revealed a normal distribution, suggesting that "merging" of probabilities is more consistent with a gradual, probabilistic process, rather than a bimodal one. A bimodal distribution would have indicated that A and B priors tended to be either perfectly merged or perfectly separated within participants (see "Merging appears graded, rather than binary, both across and within" section). These results are consistent with previous findings. For instance, Gershman & Niv, (2013) did not observe "complete" merging of the distributions across colors, either. The distributions of estimates for the two colors were significantly different from each other even when they were biased towards the mean of the two colors.

Finally, the most convincing evidence for weak fusion in our study stems from the finding that stimulus and response probabilities can be combined across cues without any loss in participants' ability to discriminate response rules across cues, and without any loss in performance at all. The "Cue noise, error rates and reaction times as a function of condition" section showed that, while the degree of merging along the response or stimulus dimensions varied with BX frequency condition, the best-fit values of the cue noise parameter (a metric of how much participants distinguish A and B response rules) did *not* vary, nor did overall error rates or reaction times. If anything, cue noise, error rates and RTs were numerically lower in the MERGE condition, which favored merging along the response dimension. In other words, participants were able to merge probabilities across cues without merging stimulus-response mappings across cues, suggesting that they were using both individual cue information and merged predictions across cues to guide behavior. Moreover, individual differences in how much a participant's best-fit A and B priors differed from each other did not correlate with cue

noise, error rates or correct RTs, providing further evidence that merging of expectations across cues operates via different mechanisms from those that maintain the response rules separate.

When A and B predict opposite responses, do participants form separate predictions for each cue?

One controversy in our results was whether, when A and B predicted opposite responses, participants nevertheless continued to group response predictions across cues. Participants' *implicit* response priors (derived from fitting their accuracy and RTs on the task) showed that, despite AX and AY frequencies remaining the same for all participants, A priors decreased across participants as BX frequency increased (as B predicted the right response more). Implicit A priors did not rise back to their original value even when B predicted the opposite response, suggesting that response predictions from the two cues remained "merged", even when they were maximally different (SPLIT condition). Given that across-participant correlations between the priors were positive overall, we found no evidence that the two parameters ever became truly independent, even in the SPLIT condition. However, participant's *explicit* reports of the trial frequencies, which we elicited at debriefing by asking them to estimate the % of each trial type (AX, AY, BX and BY) on a scale from 0 to 100, depicted a fairly different story. First, we found a non-monotonic – U-shaped – function of A priors as a function of BX/BY frequencies, in line with theories of robust estimation, which suggest that AX:AY estimates should vary with BY:BX to the degree that they suggest similar response frequencies, but cease to vary with BY:BX (and return to their original value) when the predicted response ratios grow sufficiently different. However, examining across-participant correlations between reported AX:AY ratios

and BY:BX ratios revealed that the story was even more complicated than a transition from "merging" to "splitting" response probabilities. The AX-CPT task has the interesting feature that when the cues make opposite *response* predictions, they also make identical predictions about the upcoming *stimulus*. This occurs when X probes are more frequent than Y after both cues, because A requires the left response after X while B requires the right response. Since participants are probably using the cue not only to prepare their responses, but also to predict upcoming stimuli, in the SPLIT condition, they might learn that the two cues predict the same stimulus, despite requiring opposite responses most of the time. Consistent with this idea, across-participant correlations between reported ratios revealed that, in the condition we termed "SPLIT", participants' AX estimates were positively correlated with their *BX* estimates but not significantly correlated with their BY estimates. Thus, when the two cues predicted a high probability of X, A and B estimates were positively correlated along the X/Y dimension, but not significantly correlated along the Left/Right dimension. On the other hand, in the condition we termed "MERGE", participants AX estimates were positively correlated with their *BY* estimates but not significantly with their BX estimates (**Figure 32**). Thus, when A and B predicted the same response but opposite stimuli, A and B estimates were positively correlated along the Left/Right dimension, but not significantly correlated along the X/Y dimension. A possible confound is that across-participant correlations between AX:AY and BX:BY reports could be caused by individual differences in probability learning: a participant who underestimates probability ratios in general would underestimate AX:AY and BX:BY to the same degree in the SPLIT condition, or underestimate AX:AY and BY:BX to the same degree in the MERGE condition, since those conditions differ in which pairs have the same probability ratios.

However, the fact that both AX and AY trials change their "affiliations" depending on experimental condition argues in favor of the interpretation that the way participants grouped observations across cues changed qualitatively as a function of trial frequencies.

To ensure that these correlations were not caused by individual differences in overall overestimation or underestimation of the true ratios, we performed a follow-up experiment in which we tested whether these correlations persisted even after the true probabilities had changed. Participants were assigned either to the MERGE trial frequencies (A and B predict the same response, but opposite stimuli) or to the SPLIT trial frequencies (A and B predict opposite responses, but the same stimulus) during the first 200 trials. In the second 200 trials, the frequencies of BX/BY trials changed without notice to 25%-25%, such that B no longer made any predictions about the response. We found that despite substantial carryover from the first half of the experiment (B remained left-biased for MERGE participants and right-biased for SPLIT participants), participants ultimately adapted to the change, such that A and B response priors diverged for MERGE participants and came closer together for SPLIT participants. We also examined correlations between the priors within participants across blocks, and found that participants had learned opposite ways of grouping A and B response priors in the two conditions. Implicit A and B priors were positively correlated for MERGE participants, such that an increase in left bias on A was accompanied by increase in left bias on B. SPLIT participants exhibited negative correlations, such that an increase in left bias on A was accompanied by an increase in right bias on B. Importantly, participants continued to exhibit these correlations for several blocks after the change in trial frequencies, despite the fact that the new frequencies did not exhibit any correlations between A and B. In fact, B did not predict either stimulus or

either response in Part 2 of the experiment. In other words, for the exact same trial frequencies in Part 2, MERGE participants exhibited positive correlations while SPLIT participants exhibited negative correlations between A and B response priors.

The fact that both sets of participants showed adjustment in the *values* of their response priors after the change in frequencies, but still continued to group the priors in the way they learned earlier in the task, may be further evidence of weak (rather than strong) fusion. It suggests that, even in the MERGE group, A and B response priors may have remained separate internally, since the two parameters were adjusted separately in Part 2, even though they were indistinguishable from behavior in Part 1 of the experiment.

Moreover, it is possible that our results are relevant to the distinction between "structural learning" and "parametric learning" (Braun, Mehring, & Wolpert, 2010), terms that capture two types of learning problems that biological organisms must simultaneously resolve. Parametric learning involves learning the specific parameters that maximize the performance of a model of the world, while structural learning involves learning which model of the world is relevant in the first place. The fact that participants' A and B prior values adapted soon after the change in trial frequencies, while the way in which the two parameters were grouped within participants appeared to persist until block 10 of the experiment, suggests that participants might have been adjusting the parameters of the model they had learned, until they finally gained evidence (perhaps by block 10) that the model itself had changed slightly. However, to fully substantiate these interpretations, a mechanistic model of structure learning on this task might need to be built.

Two mysteries remain about the nature of merging and splitting cue predictions in this task.

Are probabilities merged across cues even when A provides no information about B?

First, if participants tend to group response probabilities when BY>BX (MERGE conditions) and group stimulus probabilities when BX>BY (SPLIT conditions), then what happens in the middle condition (BX=BY), when B predicts a uniform distribution over both responses and stimuli? Our simulations of an agent that averages response probabilities across cues showed that A was at its lowest value in this condition, because it was the last condition for which response probabilities were similar enough across cues to be grouped (**Figure 33**, top left panel). However, A was also at its lowest in this condition for agent that grouped stimulus probabilities across cues (**Figure 33**, top right panel). Since this condition does not predispose participants towards either type of grouping, is it possible that this is the only condition in which A and B predictions are truly distinct? Or are participants divided into those who group along the stimulus and those who group along the response dimension? Across-participant correlations in AX:AY and BY:BX reports showed that this was the only condition without a significant correlation between the explicitly reported A and B ratios (**Figure 23**). However, it is possible that the correlation is not visible across participants because some participants choose to group along stimulus while others choose to group along the response dimension. Moreover, the fact that the average reported AX:AY ratio in this condition is lower than all the others (due to the U-shaped curve in the explicit estimates) suggests that participants in this condition might indeed be receiving pressure to merge along both dimensions (regardless of whether participants merge the response or stimulus probabilities across cues, AX-Left would get pulled

down – either down towards the probability of BX (25%) or down towards the probability of BY (also 25%)). Finally, if participants are indeed merging across cues in this condition, it could explain the trend we saw, particularly in Lab 1 participants, towards worse performance in this condition – higher cue noise, error rates and slower correct response times (**Figure 37**). This condition might be the one in which merging has the least advantage, since the A cue predictions really do not provide any information about the probabilities for B. In the MERGE conditions, merging response probabilities is optimal, since the two cues really do make similar response predictions. In the SPLIT condition, merging stimulus probabilities is optimal, since the cues make identical stimulus predictions and this pushes the response priors away from each other in the opposite direction. However, in the middle condition, merging either of those dimensions lower the left prior on A – away from its optimal value – and either lowers or increases the left prior on B, again away from its optimal value. **Figure 33** shows clearly that either method of merging causes the greatest probability distortions in this middle condition. Thus, most of the evidence points towards participants merging predictions in this condition, despite the fact that it may be suboptimal.

To shed further light on behavior in this condition, we plan to run the follow-up experiment design with participants who start in the BX=25% BY=25% condition and then change to 25%-25%-25%-25% (entirely uniform trial frequencies). If some participants learn to merge stimulus predictions while others learn to merge response predictions, then we should see that within-participant correlations display a bimodality, with some participants exhibiting negative and others exhibiting positive correlations within participants, rather than a normal distribution of correlations centered on 0. Second, if we see such bimodality, then participants who start out

with positive vs. negative correlations between A and B priors should continue to group them in this way after the change in frequencies. On the other hand, if participants tend not to learn any consistent grouping in this condition, then their initial within-participant correlations should not predict their subsequent correlations after the change in frequencies. Finally, participants who do not merge A and B predictions (who display null correlations between them across blocks) should actually perform better, earning higher reward rate, since grouping is suboptimal in this condition.

Why did implicit priors exhibit merging of response priors regardless of similarity, while explicit priors exhibited merging only to the degree that the cues made similar response predictions?

The second mystery that remains is the following: why did explicit priors reveal a clear non-monotonic influence of B frequencies on the A prior, while implicit priors resembled an exponential decay of A as a function of B? As shown in **Figure 33**, the U-shaped curve is consistent with an agent that either stops averaging A and B response predictions when they become dissimilar, or with an agent that begins averaging stimulus predictions instead. On the other hand, as shown in **Figure 16**, the exponential decay is more consistent with an agent that continues to weakly merge response predictions across cues, even when they are maximally different. Across-participant correlations between A and B priors also showed this mysterious discrepancy: in the implicit measure, they remained weakly positive even when the cues predicted different responses; in the explicit measure, the correlations change from positive to negative, and suggested merging of stimulus predictions when response predictions diverge.

Finally, in the SPLIT condition, the follow-up experiment showed non-significant or weakly positive correlations between A and B *across* participants, but significantly negative correlations *within* participants.

One possible explanation for these discrepancies between explicit and implicit priors, as well as between-participant vs. within-participant correlations, is that implicitly measured priors are confounded with handedness bias (the tendency of a participant to prefer one button over the other). Handedness bias is a very complex confound to deal with, since it interacts with our experimental manipulation: in the MERGE condition where A and B predict the same response, participants *should* develop a bias for one of the responses over the other, if they are correctly learning the task statistics. On the other hand, in the SPLIT condition, where left is more common for A but right is more common for B, developing separate response biases for A and B is important for task performance. In that condition, a handedness bias is suboptimal, but could nevertheless exist, if participants start out with unequal preferences for the two response buttons. Participants might, for instance, prefer the right button over the left, either because of predispositions (right-handedness), or because they strategically decide to use one of the buttons as their default, to simplify learning of task rules. Such a preference for the right response button would accelerate right responses both after A and B cues, making it appear as though A and B priors are correlated across participants (so long as there is variance across participants in which button they prefer and variance in whether they have a button preference at all.) However, handedness would not contaminate within-participant correlations across blocks, as long as it is constant for an individual participant, explaining why we can detect negative correlations in the SPLIT condition within participants, but not across participants.

Moreover, handedness differences should not affect participant's explicit reports of the probabilities at debriefing, which might explain why we can detect a U-shaped curve of A as a function of B in the explicit, but not the implicit, priors. This would also explain why a participant's explicit reports significantly predicted their implicit priors on the task, despite the fact that these two measures showed different patterns of response prior merging (i.e., the patterns in implicit prior might result from a combination of the same patterns as in explicit reports, in addition to handedness biases). It is possible that adding a handedness parameter into our models might enable us to distinguish handedness from probability learning, but the two are almost impossible to distinguish in the MERGE condition, where probability learning itself *causes* handedness. Overall, these results suggest that the within-participant correlations we performed in the follow-up experiment, as well as analyzing explicit reports, might be cleaner measures of how participants are merging probabilities across cues.

In the second part of the discussion, we comment further on the relationship between the implicit and explicit measures, as well as people's overall tendency to underestimate probability ratios, which we found in both measures.

Implicit response priors on the task were significantly correlated with explicit probability estimates at debriefing

One of our most surprising findings was that, despite the different patterns of merging between implicit priors and explicit reports, explicitly reported probabilities nevertheless predicted implicit priors on the task on a participant-by-participant basis (Figure 29). This

suggested that the priors they had learned over the course of the task, which had governed their accuracy and reaction times, at least partially influenced the probabilities they reported explicitly at debriefing. The correspondence was particularly striking in the M-turk group, where feeding participants' estimated trial frequencies into the model entirely accounted for the deviation of their starting points from optimality relative to the true frequencies (Figure 31). Indeed, both of the tendencies that had caused best-fit starting points to deviate from optimal ones were also present in their explicit reports. First, the degree of merging response probabilities across cues was similar across both measures. Although explicit reports showed a clearer separation of A and B response priors when the cues made different predictions (a clearer U-shaped curve of A as a function of B), implicit priors on M-turk still showed hints of this non-monotonic relationship, with a numerical rise in A as cue predictions diverged, suggesting that the way in which A priors varied with BX frequency was not greatly different between implicit and explicit priors on M-turk. Second, the overall tendency for starting points to be too close to 0 was equally present in explicit reports, which showed that BY:BX ratios were underestimated when they were high, and overestimated when they were low ("Implicit and Explicit Priors are Closer to 0 than Optimal" section).

Unlike M-turk participants, lab participants' explicit estimates did not account for the sub-optimality in the starting points they used on the task (**Figure 31**), despite the significant correlation between the two measures participant-by-participant. The two deviations from optimality that were present in M-turk participants' implicit priors were also present in those of lab participants, but were not sufficiently present in the explicit reports of lab participants. First, the degree of merging along the response dimension was much greater in lab participants'

implicit priors than their explicit ones: implicit A priors continued to decrease as a function of BX frequency, suggesting that B was exerting an influence on A even when their predictions diverged, whereas explicit A priors were not pulled down by B as much and rose back to their original level when the cues made different predictions. Thus, for lab participants, response priors were merged much more in implicit priors than in the explicit reports. Second, we did not find significant underestimation of probability ratios in lab participants' explicit reports: the curve of estimated BY:BX ratios closely followed the identity line, and all deviations from the identity line could be explained by B being pulled towards A (merging across cues), rather than ratios being pulled towards 50/50. This was different from M-turk participants' explicit reports, where the estimated BY:BX ratios had a significantly shallower slope than the identity line.

To summarize, lab participants' explicit probability estimates were more accurate – closer to the true probabilities – than those of M-turk participants. They showed less merging of response priors across cues, and less underestimation of probability ratios. However, since lab participants' implicit priors on the task showed the same deviations from optimality as those of M-turk participants, this caused a discrepancy between the explicit probability estimates and their implicit priors.

This raises two questions. First, why do lab participants show substantially less underestimation of probability ratios in their explicit estimates? Second, if they can report such accurate probabilities explicitly, why is this learning not implemented in their behavior on the task? One possibility is that lab participants are less distracted during performance of the task, or have more cognitive resources at their disposal (possibly due to higher working memory capacity), such that they are more likely to notice the unequal probabilities on the task. They

might begin tracking these probabilities deliberately across trials, forming more accurate counts of each trial type. Such an interpretation would suggest that explicit knowledge of the probabilities does not necessarily counteract the biases in people's priors on the task. Future experiments could elucidate the relationship between implicit priors and explicit estimates by asking M-turk participants to deliberately track trial frequencies during the instructions phase, and to reward them for the accuracy of their estimates at debriefing. If deliberately tracking probabilities corrects the biases towards smaller ratios in their explicit reports but not in their behavior during the task, it would suggest that different mechanisms are driving implicit priors vs. explicit estimates. Participants could even be told the true frequencies from the very beginning, to verify how this explicit knowledge affects performance.

Despite these remaining questions, our existing data already constrains the mechanisms by which merging of probabilities across cues could operate. Specifically, comparing implicit and explicit reports sheds light on whether merging arises from distortions in *representation* vs. *implementation* of priors. One extreme scenario is that participants could accurately learn probabilities across trials, forming entirely separate predictions for A and B cues, but that merging arises during the implementation of these priors on individual trials. For instance, during the delay period between cue and probe, participants might sample from their memories of both A and B trials, settling on an effective prior that is skewed towards the real cue on that trial, but also biased towards the average between the two. The second extreme scenario is that merging operates on the level of trial-by-trial probability learning, such that participants' memories of A and B priors are already distorted towards the average between them, but that the implementation on individual trials is a perfect reflection of those memories.

The fact that explicit reports show merging across cues in all participant groups suggests that merging does not arise solely from sub-optimality during implementation and that it affects the representations of probabilities that participants learn. In M-turk participants especially, the fact that starting points were optimal given their probability estimates (when accounting for overly high decision thresholds) suggests that both distortions (both merging across cues and underestimation of probability ratios) operate at the level of probability learning, but that the implementation of the learned priors on individual trials is accurate. Of course, the fact that lab participants' priors are sub-optimal relative to their frequency estimates does not necessarily prove that the sub-optimality arises during implementation on individual trials. It is also possible that their explicit reports access a different representation from the one used during the task to generate behavior. However, the close correspondence between the implicit and explicit measures in M-turk participants makes it likely that both measures are accessing the same representation of learned probabilities, and argues in favor of distortions in *representation* rather than distortions in *implementation* of priors. This distinction between perfectly separated memories of A and B probabilities that are merged on individual trials and distorted memories that are perfectly implemented on individual trials becomes important in the next chapter of this dissertation. In **Study 3**, we will show through model simulations that merging of starting points across A and B cues can generate choice behavior that has been taken as a index of reactive control strategies in the AX-CPT (Braver, 2007). In the literature, such behavior has been interpreted to mean that participants may not be representing the cue in working memory during the delay period between cue and probe (Braver, Paxton, Locke, & Barch, 2009). Based on the findings above, we will argue that merged starting points between A

and B can arise from distortions in probability learning over the course of hundreds of trials, as opposed to sub-optimal implementation of distinct A and B priors on a single trial level.

Response priors were lower than optimal and probability ratios were underestimated

By comparing participants' best-fit starting points with their probability estimates, our study also adds to the literature on drift diffusion models of optimal decision-making in two-alternative force choice tasks (Bogacz et al., 2006). Several studies have investigated whether human participants adopt the thresholds that maximize reward rate in simpler (context-free) tasks with only two choices, and found that most participants use thresholds that are higher than optimal, possibly because they prioritize accuracy in addition to reward rate, or because they misestimate durations (Balci et al., 2011; Bogacz, Hu, Holmes, & Cohen, 2010; Zacksenhouse, Bogacz, & Holmes, 2010). As far as we are aware, a smaller number of studies have investigated how participants' starting points vary as the probability of the two choices varies, and whether they deviate from optimal ones (Leite & Ratcliff, 2011; Ratcliff & McKoon, 2008; Simen et al., 2009). Leite and Ratcliff (2011) varied the probability of correct low vs. high responses for a task in which participants had to judge whether the number of asterisks on screen was higher or lower than some criterion. They found that varying the probability ratio of high to low responses between 1:3, 1:1 and 3:1 affected participants' starting point but did not affect drift rate. However, they did not assess whether participants' best-fit starting points maximized reward rate, and three ratios is not sufficient to determine the functional relationship between participants' starting points and the true probability ratios. Simen et al.

(2009) aimed to verify whether participants' adapted their starting points optimally to maximize reward rate in response to changes in stimulus frequencies. Participants had to evaluate the direction of dot motion (whether dots were moving to the left or right) and the probability of left vs. right motion varied across blocks. Probability ratios of 50:50, 60:40, 75:25 and 90:10 were tested (the identity of left – right was counterbalanced, as in our study). They found that the best-fit starting points varied monotonically (possibly linearly) with the probability ratios, but the best-fit starting points were consistently below optimal ones, such that the best-fit starting point curve was much shallower than the optimal curve, exactly as we found in our study. Simen et al. (2009) were uncertain why participants did not adapt their starting points to the probability ratios as strongly as the optimal model predicted. Our study suggests that overly low starting points might arise directly from distortions in the way participants learn probabilities, as opposed to sub-optimal decision-making. (At least for M-turk participants, overly low starting points could be entirely explained by their distorted probability estimates.)

Possible explanations for underestimation of probability ratios

Though this phenomenon has not been the focus of studies testing optimality predictions of drift diffusion models, the finding that people underestimate probability ratios has been very extensively replicated in the literature on the estimation and use of probabilities in decisions across a broad range of domains and paradigms (Zhang & Maloney, 2012). For instance, Tversky & Kahneman (1992) and Gonzalez & Wu (1999) gave participants a choice between a certain payoff and a gamble with known probabilities of winning and losing. Based on participants'

choices, it appeared that they underestimated high probabilities of winning and overestimated low probabilities. Participants display similar biases when estimating the frequencies of life events (Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978) or words in the English language (Attneave, 1953). In the domain of perceptual decisions, people display similar biases when estimating the proportion of white vs. black dots on a screen (Varey, Mellers, & Birnbaum, 1990). Finally, studies testing signal detection theory have found that as the prior probability of a signal varies (such as the probability of an auditory signal), participants appear to adjust their decision criterion in suboptimal ways, overestimating small probabilities and underestimating large ones (Green & Swets, 1966). For all of the above datasets, Zhang and Maloney (2012) found that estimated probabilities followed an S-shaped curve as a function of true probabilities, and that such S-shaped distortions were well-captured by the following family of functions:

$$\log\frac{p'}{1-p'} = \gamma \log\frac{p}{1-p} + (1-\gamma)\log\frac{p_0}{1-p_0}$$

In our notation, p is the true probability, p' is the estimated probability, and p_0 is the fixed point of the transformation, the value of p that is perfectly estimated without distortion. These functions appear linear when the log ratios of the probabilities are plotted, with γ being the slope of the line. If $\gamma < 1$, the curve of estimated probabilities will have a shallower slope than the true probabilities, such that small probabilities are underestimated while large ones are overestimated (as is the case in our study, as well as all the studies cited above). The p_0 is the value for which the estimated ratio corresponds to the true ratio, which in our study tended to be 50%-50% (see **Figure 20**, right panel). Zhang and Maloney (2012) experimentally

manipulated the number of possible options, and found that p_0 changed from close to 0.5 for two options to approximately 0.25 for four options, suggesting that p_0 may be the prior probability of each option if the probability distribution is uniform (people start out believing that all choices are equally likely).

Despite the ubiquitous nature of this finding, very little is understood about the mechanisms that drive probability ratio underestimation across such a wide range of paradigms (Zhang & Maloney, 2012). Importantly, Glaser, Trommershauser, Mamassian, & Maloney (2012) found that underestimation of probability ratios persisted even when participants were given hundreds of trials to experience the probabilities of events, suggesting that it does not merely result from participants beginning with a uniform prior and needing more data before their prior can be overridden. These results are consistent with our findings in Experiment 2: fitting the DDM block-by-block showed that best-fit A and B priors appeared to asymptote after only a few blocks (**Figure 25**), rather than continually increasing across blocks (which would have suggested that they might reach optimal values with sufficient experience). In fact, our Lab 1 participants in Experiment 1 received twice as many trials (480) as the Lab 2 and M-turk groups (240), yet their implicit priors remained smaller than optimal, and the degree of underestimation did not differ significantly between Lab 1 and Lab 2, suggesting that having twice as much experience did not enable Lab 1 participants' priors to approach the values that maximize reward rate (**Figure 31**). Thus, underestimation of probability ratios does not appear to arise merely from uncertainty caused by insufficient experience with the probabilities.

It is interesting to note that Rescorla-Wagner models (Rescorla & Wagner, 1972), which describe how associations between cues and stimuli may be learned incrementally across trials

via prediction errors, can only give rise to such underestimation when their learning rate is very low, such that 480 trials does not provide the model with enough trials to learn the probability ratios, resulting in estimates that are closer to the model's 50-50 prior. However, given enough trials, the model ultimately learns probability ratios that are very close to the true numbers. In fact, with higher learning rates, such models exhibit the opposite behavior, underestimating low probabilities and overestimating high ones, due to the exponentially decaying influence of past experience on current estimates (this is because recent experience has a higher influence on estimates, and recent trials are more likely to contain high-probability than low-probability events; Niv, Joel, Meilijson, & Ruppin, 2002). This suggests that Rescorla-Wagner models cannot account for probability learning in our study. Future studies could explore whether a Pearce-Hall model (Pearce & Hall, 1980), in which large prediction errors result in greater learning than small ones, could better explain the overestimation of low probabilities (low-probability outcomes in such models could result in greater surprise and greater learning than high-probability outcomes, which might result in closer probability estimates for the two outcomes). Models with learning rates that change with uncertainty (Mathys, Daunizeau, Friston, & Stephan, 2011), as well as models with multiple learning systems (Bornstein & Daw, 2012, 2013), could also provide insights into this data.

All in all, the explicit reports of M-turk participants, which exhibit the S-shaped distortion, are more consistent with the rest of the literature than the explicit reports of lab participants, which appear remarkably close to the identity line. Only one previous paradigm we are aware of has reported remarkably accurate probability estimates that did not deviate from the identity line (Gallistel, Krishan, Liu, Miller, & Latham, 2014; Robinson, 1964). On every trial,

participants witnessed a single sample from an urn that had some fixed probability of producing green vs. red circles and adjusted a slider to indicate their perceived probability of green circles in the box. The probability tended to remain stable for long periods of time, and jump to a different value unexpectedly (participants were informed that the box might be "replaced" randomly from time to time and were asked to indicate their belief that a change had occurred). Gallistel et al. (2014) found that participants' estimates in such experiments did not deviate from the true probabilities. One possible explanation for their accuracy is that participants were instructed to report their estimates of probability, such that they could compare their predictions with the outcome of the draw on every trial. In our study, as well as that of Glaser et al. (2012), participants' expectations were measured indirectly from their choices (implicit measure), such that participants' attention was not drawn to the task of estimating probabilities explicitly. It is possible that instructing participants to focus on probabilities in this way, enabling them to compare their predictions with outcomes, produces highly accurate reports. Moreover, it is possible that encouraging participants to indicate whether the urn generating their observations has changed unexpectedly alerted them to the fact that the urn actually remained the same across trials and only changed at exceptional "change points". In contrast, our participants did not know anything about the underlying structure of the task, and specifically did not know whether the probabilities would remain stable over time. As such, they might have employed learning algorithms that attempt to simultaneously infer the current probabilities and whether the underlying model of the world has changed (Iglesias et al., 2013; Mathys et al., 2014; Nassar, Wilson, Heasly, & Gold, 2010; Wilson, Nassar, & Gold, 2013). Such algorithms may result in greater adaptability in constantly

changing environments, but at the expense of less accurate probability estimates when the environment is stable.

Based on Gallistel et al. (2014), it seems possible that lab participants provided more accurate explicit estimates because greater cognitive flexibility enabled them to reach explicit awareness of the unequal probabilities in the task. It is possible that such explicit estimation does not influence behavior during the task, which would explain the discrepancy between implicit and explicit priors in this group. If this is the case, it would suggest that the strong correspondence between implicit and explicit priors on M-turk arises from lack of explicit estimation of probabilities during the task in M-turk participants, such that explicit reports at debriefing must rely on implicitly learned representations.

Conclusion

We investigated how varying stimulus probabilities given a contextual cue would affect the degree to which people used contextual information in their behavior, both in terms of adapting their stimulus-response rules based on the cue and in terms of using the contextual cue to prepare the most likely response. Based on previous findings from perceptual estimation, we hypothesized that when two contexts predict similar response probabilities, people should "merge" probabilities across contexts, and that the degree of merging should be proportional to the similarity of contextual predictions. Confirming our hypothesis, we found that people's expectation of the response (as measured by their accuracy and reaction times) was closer to the average probabilities across the two contexts, but only to the degree that the

two cues made similar response predictions. Moreover, when the cues made different response predictions but similar stimulus predictions, participants also appeared to merge stimulus probabilities across cues. Importantly, merging probabilities across cues did not come at the expense of merging cue representations: participants were able to distinguish stimulus-response mappings by cue equally well regardless of how much they merged probabilities. Examining their explicit reports of stimulus probabilities at debriefing revealed that stimulus and response probabilities were merged across cues in these reports as well, suggesting that merging did not arise from incorrect implementation within trials, but appeared to stem from distorted probability learning across trials. In online participants only, response priors were indistinguishable from optimal given their reported probabilities. Finally, a follow-up experiment showed that participants who had learned to group response probabilities across cues continued to do so even when the trial frequencies changed, such that the environment no longer presented any similarities between cue predictions. Analogously, participants who had learned to group stimulus probabilities across cues continued to do so after a change in trial frequencies, suggesting that the similarity structure that participants experienced at the beginning of the task influenced the way they grouped observations across cues later in the experiment.

Prior Presentations, Publications and Co-Authorship

This work was completed in collaboration with Robert C. Wilson, Michael Shvartsman and Jonathan D. Cohen. Initial stages of the project have been presented as posters at two conferences:

Lositsky, O., Wilson, R.C., Shvartsman, M., Cohen, J.D. (2015, December). Adaptive task representations in context-based decision making. "Bounded Optimality and Rational Metareasoning" workshop, *Neural Information Processing Systems Conference*, Montreal, Canada. [***Poster***]

Lositsky, O., Wilson, R.C., Shvartsman, M., Cohen, J.D. (2016, November). Adaptive task representations in context-based decision making. *Society for Neuroscience Annual Meeting*, San Diego, CA. [***Poster***]

The work is also forecast to be presented at CogSci 2017 and SfN 2017:

Lositsky, O., Shvartsman, M., Wilson, R.C., Cohen, J.D. (2017, July). Adaptive response priors in context-dependent decision-making. *39th Annual Meeting of the Cognitive Science Society*, London, UK. [***Poster***]

Lositsky, O., Shvartsman, M., Cohen, J.D., Wilson, R.C. (2017, November). Contributions of working memory and across-trial probability learning to context-based decisions. *Society for Neuroscience Annual Meeting, Washington, DC.* [***Poster***]

References

Attneave, F. (1953). Psychological Probability as a Function of Experienced Frequency. *Journal of Experimental Psychology*, *46*(2). doi:http://dx.doi.org/10.1037/h0057955

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics*, *73*(2), 640–657. doi:10.3758/s13414-010-0049-7

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–65. doi:10.1037/0033-295X.113.4.700

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do humans produce the speed-accuracy trade-off that maximizes reward rate? *Quarterly Journal of Experimental Psychology (2006)*, *63*(5), 863–891. doi:10.1080/17470210903091643

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, *35*(7), 1011–1023. doi:10.1111/j.1460-9568.2011.07920.x

Bornstein, A. M., & Daw, N. D. (2013). Cortical and Hippocampal Correlates of Deliberation during Model-Based Decisions for Rewards in Humans. *PLoS Computational Biology*, *9*(12). doi:10.1371/journal.pcbi.1003387

Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural Brain Research*. doi:10.1016/j.bbr.2009.08.031

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7351–6. doi:10.1073/pnas.0808187106

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. doi:10.1038/415429a

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2004.02.002

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*. doi:10.1037/a0035232

Genewein, T., & Braun, D. A. (2014). Occam's Razor in sensorimotor learning. *Proceedings. Biological Sciences / The Royal Society*, *281*, 20132952. doi:10.1098/rspb.2013.2952

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics

combine in size perception. *Current Biology*, *13*(6), 483–488. doi:10.1016/S0960-9822(03)00133-7

Gershman, S. J., & Niv, Y. (2013). Perceptual estimation obeys Occam's razor. *Frontiers in Psychology*, *4*(SEP). doi:10.3389/fpsyg.2013.00623

Girshick, A. R., & Banks, M. S. (2009). Probabilistic combination of slant information: weighted averaging and robustness as optimal percepts. *Journal of Vision*, *9*(9), 8.1-20. doi:10.1167/9.9.8

Glaser, C., Trommershauser, J., Mamassian, P., & Maloney, L. T. (2012). Comparison of the distortion of probability information in decision under risk and an equivalent visual task. *Psychol Sci*, *23*(4), 419–426. doi:10.1177/0956797611429798

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166. doi:10.1006/cogp.1998.0710

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *Society*, *1*, 521. doi:10.1901/jeab.1969.12-475

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science (New York, N.Y.)*, *298*(5598), 1627–1630. doi:10.1126/science.1075396

Hospedales, T., & Vijayakumar, S. (2009). Multisensory oddity detection as bayesian inference. *PLoS ONE*, *4*(1). doi:10.1371/journal.pone.0004205

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., denOuden, H. E. M., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530. doi:10.1016/j.neuron.2013.09.009

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching?A review. *Psychological Bulletin*, *136*(5), 849–874. doi:10.1037/a0019842

Knill, D. C. (2007). Robust cue integration : A Bayesian model and evidence from cue-con fl ict studies with stereoscopic and fi gure cues to slant. *Journal of Vision*, *7*(5), 1–24. doi:10.1167/7.7.5.Introduction

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, *35*(3), 389–412. doi:10.1016/0042-6989(94)00176-M

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making*, *6*(7), 651–687. Retrieved from http://ideas.repec.org/a/jdm/journl/v6y2011i7p651-687.html

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology Human Learning and Memory*, *4*(6), 551–578. doi:10.1037/0278-7393.4.6.551

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*(November), 825. doi:10.3389/fnhum.2014.00825

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39. doi:10.3389/fnhum.2011.00039

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(37), 12366–78. doi:10.1523/JNEUROSCI.0822-10.2010

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviors. *Adaptive Behavior*, *10*(1), 5–24. doi:10.1177/10597123020101001

Pacer, M., & Lombrozo, T. (2017). Ockham's Razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology:General*.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. doi:10.1037/0033-295X.87.6.532

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi:10.1162/neco.2008.12-06-420

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*(6), 64–99. doi:10.1101/gr.110528.110

Robinson, G. H. (1964). Continuous estimation of a time-varying probability. *Ergonomics*, *7*(1), 7–21. doi:10.1080/00140136408930721

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1865–1897. doi:10.1037/a0016926

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of

uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. doi:10.1007/BF00122574

Varey, C. a, Mellers, B. a, & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology. Human Perception and Performance*, *16*(3), 613–625. doi:10.1037/0096-1523.16.3.613

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2013). A Mixture of Delta-Rules Approximation to Bayesian Inference in Change-Point Problems. *PLoS Computational Biology*, *9*(7). doi:10.1371/journal.pcbi.1003150

Zacksenhouse, M., Bogacz, R., & Holmes, P. (2010). Robust versus optimal strategies for two-alternative forced choice tasks. *Journal of Mathematical Psychology*, *54*(2), 230–246. doi:10.1016/j.jmp.2009.12.004

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, (JAN). doi:10.3389/fnins.2012.00001

Study 3: Distinguishing Proactive and Reactive Control Strategies in Behavior using a Drift Diffusion Model

Abstract

Most of our everyday decisions rely crucially on context: foraging for food in the fridge may be appropriate at home, but not at someone else's house. Yet the mechanism by which context modulates how we respond to stimuli remains a topic of intense investigation. In order to isolate such decisions experimentally, investigators have employed simple context-based decision-making tasks like the AX-Continuous Performance Test (AX-CPT; Rosvold, Mirsky, Sarason, Bransome, & Beck, 1956). In this task, the correct response to a probe stimulus depends on a cue stimulus that appeared several seconds earlier. It has been proposed (Braver, Gray, & Burgess, 2007) that humans might employ two strategies to perform this task: one in which rule information is *proactively* maintained in working memory, and another one in which rule information is retrieved *reactively* at the time of probe onset. While this framework has inspired considerable investigation, it has not yet been committed to a formal model. Such a model would be valuable for testing quantitative predictions about the influence of proactive and reactive strategies on choice and reaction time behavior. To this end, we have built a drift diffusion model of behavior on the AX-CPT, in which evidence accumulation about a stimulus is modulated by context. We implemented proactive and reactive strategies as two distinct

variants: in the proactive variant, perception of the probe is modulated by the remembered cue; in the reactive variant, retrieval of the cue from memory is modulated by the perceived probe. Simulations from the two variants reveal that the proactive model can produce reactive choice behavior under realistic settings of the parameters, but cannot produce reactive reaction time behavior. Moreover, fitting these models to data shows that, counter-intuitively, behavior taken as a signature of reactive control is better fit by the proactive variant of the model, while proactive profiles of behavior are better fit by the reactive variant. We offer possible interpretations of this result, and use simulations to suggest experimental manipulations for which the two models make divergent predictions.

Introduction

Most of our daily decisions depend crucially on context. While our normal response to seeing a new email from a good friend might be to open it, this response may be suppressed if our goal is to work. However, if an email alert were to arrive about an approaching winter storm, we should be able to respond to it despite its former irrelevance. This ability to modulate our behavior based on context has been termed "cognitive control." The traditional account of these context effects is a *proactive* one, in which context is represented by patterns of ongoing neural activity (possibly sustained by the frontoparietal network) and modulates the response to incoming stimuli (Miller & Cohen, 2001). Recently, the Dual Mechanisms of Control (DMC) account has proposed that, since active task preparation is costly (the preparation time could be used for a different purpose), people may sometimes use a *reactive* strategy. Instead of actively maintaining the context, it could be stored passively (perhaps in a pattern of synaptic weights) and reactivated when the relevant stimulus is presented (Braver et al., 2007). Under the DMC framework, people can switch between these two strategies based on their intrinsic motivation, the constraints of their cognitive architecture, and the particular demands of the task (Braver et al., 2007).

The DMC framework has inspired a large body of experimental work using the AX-Continuous Performance Test (AX-CPT) in humans (see Braver (2012) for a review). In this task, a contextual cue (A or B) is presented briefly on each trial, followed by a probe stimulus (X or Y) after a short delay. Participants must respond to the probe stimulus differently depending on whether it was preceded by (i.e. in the context of) an A or a B. The AX-CPT has been used

extensively to diagnose disorders of cognitive control in various patient populations (Rosvold et al., 1956; see Riccio, Reynolds, Lowe, & Moore (2002) and Bellani & Brambilla (2008) for reviews). In this study, we provide the first computational instantiation of dual mechanisms of control in the AX-CPT[4].

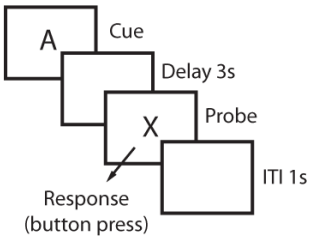
We model proactive control as identifying the probe (X or Y) in the context of the cue (A or B), and reactive control as recalling the cue (A or B) in the context of the probe (X or Y). We instantiate the probe identification or cue recall processes as drift-diffusion models that allow us to predict the pattern of choices and reaction times for the different trial types in the task. Using these models, we show that choice behavior taken as a signature of proactive or reactive control can be produced by both instantiations of our model, though they make different predictions for reaction time distributions. Moreover, we find – counter-intuitively – that proactive behavior is better fit by the model we termed "reactive", while reactive behavior is better fit by the model we termed "proactive", and offer possible interpretations. Our simulations suggest that manipulating uncertainty about the cue and probe separately could bring participants into a parameter regime where the two models make the most distinct predictions.

[4] De Pisapia & Braver (2006) built an early neural network model of dual mechanisms of control in the Stroop task. In that model instantiates, reactive control arises through sequential adjustment effects implemented by the Anterior Cingulate Cortex. However, that model's account of reactive control differs from accounts espoused by subsequent literature (Braver, 2012).

Methods

The AX-CPT Task

On each trial of the task, a cue (A or B) signals how to respond to the upcoming probe (X or Y), which appears a few seconds later. In the variant we use, the subject must press the left button on AX and BY trials, and the right button on AY and BX trials (**Figure 39**), though the keyboard buttons corresponding to left and right are counter-balanced across participants.



AX	50%	Left
AY	20%	Right
BX	20%	Right
BY	10%	Left

Figure 39 Trial timeline, trial frequencies and response rules.

Manipulating the frequencies of cue and probe stimuli can be used to investigate participants' control strategy. Since AX trials are the most common, a subject proactively preparing responses will be biased to incorrectly respond Left on AY trials, since X is more likely than Y. In contrast, a subject reacting to the probe will be biased to incorrectly respond Left on BX trials, since A precedes X more often than B. The relative proportion of AY and BX errors and reaction times in such designs have therefore been taken to reflect an index of proactive vs. reactive control (Braver, Paxton, Locke, & Barch, 2009).

Behavioral Experiments to Induce Proactive and Reactive Strategies

To encourage participants to use both types of control, we used a number of within-subject manipulations. In the proactive manipulation, participants were rewarded for correctly responding faster than 500ms, which is difficult without proactive task preparation. In the reactive manipulation, we introduced a distractor task between the cue and probe, making it more difficult to actively maintain the context during the delay period, so that participants would be encouraged to retrieve their context representation (presumably from episodic memory) at the time of the probe.

We ran 11 small experiments, varying trial frequencies (while maintaining AY and BX frequencies equal), the duration of the delay period, and the difficulty of the distractor task (in the case of the reactive manipulation). While these variations are interesting and will be reported separately, here we aggregate the results from all experiments (120 participants) in order to draw general inferences about the proactive-biased and reactive-biased conditions.

Drift Diffusion Model of Context-Dependent Decision-Making on the AX-CPT

At its heart, the AX-CPT is a two-alternative forced choice (2AFC) task: participants make a decision between left and right responses to a probe stimulus (X or Y). Such 2AFC tasks have been modeled successfully using drift diffusion models (DDMs), in which a decision is made by gradually accumulating evidence for either left or right response and responding when the evidence crosses some decision threshold, *z* (see Bogacz, Brown, Moehlis, Holmes, & Cohen,

2006, for a detailed review). Unlike classic 2-AFC tasks, however, in the AX-CPT, participants must integrate memory of the cue they saw seconds earlier (the A or the B) with incoming perceptual evidence about the probe, and the dynamics of this integration should vary by strategy.

We implemented the proactive strategy as a drift diffusion process reflecting perceptual evidence accumulation about the probe, the parameters for which are set by the noisily encoded (but instantaneously retrieved) cue in memory. We modeled the reactive strategy in complementary form: as a drift diffusion process reflecting retrieval of the cue from memory, the parameters of which are set by a noisily (but instantaneously) perceived probe. This distinction reflects the DMC conception of proactive control as task configuration triggered by the cue and reactive control as task configuration triggered by the probe.

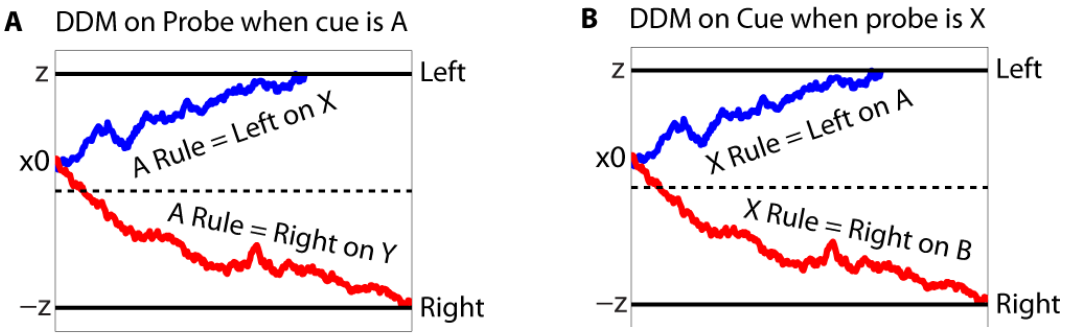


Figure 40 Proactive DDM on Probe (A) and Reactive DDM on Cue (B). In the "perfectly proactive" Drift Diffusion on Probe (DDP) model, the cue (A or B) is assumed to be entirely processed before the probe (X or Y) appears and influences the parameters for processing the probe, including the starting point (x_0), threshold (z) and direction of drift (Left on X vs. Right on X). In the "perfectly reactive" Drift Diffusion on Cue (DDC) model, the probe (X or Y) is assumed to be entirely processed before retrieval of the cue (A or B) from memory and sets the parameters for retrieving the cue.

Drift Diffusion-on-Probe (DDP) model of "purely proactive" strategy

In the proactive variant, which we will term "Drift Diffusion-on-Probe" (DDP), the true cue, *C*, is encoded in memory as $\tilde{C}$, with some bit-flip noise, ε_C, which reflects the probability that the cue is encoded incorrectly. Given its noisy cue representation, the DDP computes the posterior probability distribution over the rule, *R* (i.e. R = "left on X, right on Y" or "right on X, left on Y") using Bayesian inference:

$$P\left(R\middle|\tilde{C}\right) \propto P\left(\tilde{C}\middle|R\right)P(R)$$

with the prior set based on trial frequencies, and the likelihood computed using the true value of ε_C. The model then samples the rule from $p\left(R\middle|\tilde{C}\right)$, and uses this sampled rule to set up the parameters of the DDM on the probe. The sampled rule determines the direction of the drift, the threshold, *z*, and starting point, x_0, for drift diffusion on the probe. The model therefore has two starting point and threshold parameters (one for each cue: $\tilde{z}^A, \tilde{z}^B, {x_0}^A, {x_0}^B$), a drift rate parameter for the probe ($\tilde{a}$), cue noise (ε_C) and a response time offset due to motor planning and execution (T_0) – seven free parameters in total.

These assumptions, in combination with standard results from the DDM literature (Busemeyer & Townsend, 1993; Navarro & Fuss, 2009; Ratcliff, 1978), allow us to compute reaction time distributions for correct and error trials in each of the four trial types (AX, AY, BX, BY). This, in turn, allows us to fit the model to the experimental data.

To develop an intuition about how the model behaves, consider that the probability of using Rule B given that $\tilde{C}$=B increases with the frequency of B cues and decreases with encoding noise, ε_C. As ε_C increases, the model will be more likely to confuse B for A cues, given the

higher frequency of A. This will cause it to respond incorrectly on BX, and produce the typically "reactive" mistake. In other words, the higher the cue noise, the more reactively the "proactive" model behaves.

Drift Diffusion-on-Cue (DDC) model of "purely reactive" strategy

For the reactive variant, which we term "Drift Diffusion-on-Cue" (DDC), we reversed the order of cue and probe processing in order to capture how reactivation of task representations can be triggered and modulated by the probe.

In this variant, the probe is encoded with some bit-flip noise ε_P, and must be decoded in order to set the rule (e.g. Rule = "Right on A, Left on B") and thereby guide the response. As the probe encoding noise increases, given the higher frequency of X's, the model will be more likely to confuse Y's for X's, causing it to respond incorrectly on AY trials. Thus, mirroring the effects of the DDP, the DDC will produce typically "proactive" mistakes as probe noise increases. The sampled rule also sets the starting point for drift diffusion on the cue. Given the trial frequencies (**Figure 39**), a rational subject might set the X starting point closer to the Left, but the Y starting point closer to the Right threshold.

Note that both variants of the model implement noise in the cue and probe representations, but they do so in different ways.

In the DDP model, noise in cue memory is implemented by ε_C (inducing BX errors), whereas noise in probe perception is implemented as a low drift rate (inducing AY errors and slower AY correct responses). Conversely, in the DDC model, noise in cue memory is implemented as a

low drift rate (inducing BX errors and slower BX correct responses), whereas noise in probe perception is implemented as ε_P (inducing AY errors).

Fitting the Model to Data

We used maximum likelihood estimation to fit the DDP and DDC model variants separately for each subject in each experimental condition, and used the Bayesian Information Criterion (BIC; Schwarz, 1978) to evaluate which model fit the data best in which condition. However, BIC does not directly assess the quality of the fit. Therefore, we computed a "pseudo R^2" statistic (Fernandes, Stevenson, Phillips, Segraves, & Kording, 2014) that measures how well the models fit the data relative to a "null model" (with minimal assumptions) and a "saturated model" (that overfits the data). Following Fernandes et al. (2014), we defined the pseudo R^2 as:

$$\boldsymbol{pR^2 \stackrel{\text{def}}{=} 1 - \frac{L_S - L_T}{L_S - L_H}}$$

where $L_S$is the log likelihood of the saturated model, L_T is the log likelihood of the theoretically-informed model, and $L_H$is that of the homogeneous or null model. These benchmark models should reflect the most and least constrained forms of the theoretically-informed models. The homogeneous model assumes no variation by trial type and assumes all trials are drawn from a mixture of two arbitrary DDMs. The saturated model assumes no shared properties between trial types and assumes each trial type is drawn from a separate mixture of two arbitrary DDMs. (We make the mixture-of-DDMs assumption because our models produce a mixture of two DDMs by occasionally sampling from the wrong rule due to encoding noise.)

The resulting saturated model has 32 parameters, and thus many more degrees of freedom than the theoretically-informed models (DDP and DDC) with 7 parameters each.

A pseudo R^2 value of 0 means the theoretical model does no better than a null model that does not distinguish between trial types, and a value of 1 means that the theoretical model performs as well as a saturated model that fits each trial type separately.

Landscaping Analysis: Finding where the models make distinct predictions

To infer in which part of the parameter space the DDP and DDC models make the most distinct predictions, we performed a "landscaping analysis" (Navarro, Pitt, & Myung, 2004): we simulated behavior with each model, sampling uniformly from the parameter space, and fit this behavior with both models (repeating with both models as the simulating model). This analysis lets us identify parts of the parameter space where one model cannot fit the other's generated data, and therefore the models make distinct predictions.

Given the number of experimental design parameters available for a task even as simple as the AX-CPT, it is possible that a particular experimental design (i.e., one that implements a particular set of task parameters) may not be maximally sensitive to differences in predictions made by two relatively similar models (such as the DDP and DDC). To address this issue, and as an aid in interpreting our results, we used a "landscaping analysis" (Navarro et al., 2004) to evaluate the sensitivity of different task designs (i.e., ones defined by different model parameter sets) in distinguishing between the two models. Specifically, we used each model in turn as the "simulating model", to simulate behavior, by sampling uniformly from a space of task-relevant parameters: starting points (${x_0}^A, {x_0}^B$), thresholds (z^A, z^B), drift rate ($\tilde{a}$) and

encoding noise (ε_C for DDP and ε_P for DDC). We then evaluated the fit of each model to the behavior generated by the simulating model. We used this procedure to characterize, for each part of the parameter space, the extent to which each model could fit the other's generated data, thus identifying parameter domains where the models make similar vs. distinct predictions.

Results

Model Simulations

We first tested how varying the DDP model parameters impacted the behavioral index of proactive control, as it has been defined by previous literature. According to the definition in Braver et al. (2009) and Gonthier et al. (2016), we defined the behavioral index as:

$$ProactiveIndex = \frac{AY - BX}{AY + BX}$$

For error rates, this signified subtracting BX from AY error rates, and for reaction times (RTs), this signified subtracting mean correct BX RTs from mean correct AY RTs. In both cases, a positive proactive index suggests worse AY than BX performance and has been interpreted as more proactive behavior, whereas a negative proactive index suggests worse BX than AY performance, and has been interpreted as more reactive behavior.

Through simulations, we calculated the proactive choice and RT index for each combination of the below model parameter values:

$$\frac{cueNoise}{driftRate} = [0.005 : 0.005 : 0.05]$$

$$drift\ rates\ a = [0.5, 1, 2, 3]$$

$$thresholds\ z = [0.5, 1, 2, 3]$$

$$\frac{{x_0}^A}{z} = [0 : 0.05 : 0.2]$$

$$\frac{{x_0}^B}{z} = [-1 : 0.05 : 1]$$

$$Trial\ Frequencies = [50\%\ 20\%\ 20\%\ 10\%]$$

In each simulation, the threshold was held at the same value for both cues ($z^A = z^B$), whereas the relative distances from the starting points ${x_0}^A$ and ${x_0}^B$ to the threshold varied. The starting point parameter after A cues (${x_0}^A$) was only tested with values above 0 (closer to the left than the right boundary), since in all experimental designs used in the literature, the frequency of left responses after A was higher than the frequency of right responses and the left response was more common overall. In contrast, both positive and negative values were tested for the starting point after B cues (${x_0}^B$): even though the frequency of right responses was higher after B cues, the overall frequency of left responses in the task was higher, and participants' starting points after B cues might be influenced by both of these factors.

The purely proactive DDP model can generate reactive choice behavior

All the varied parameters significantly impacted the proactive error rate index (i.e., proactive choice index). Due to the large number of simulations, all p-values were indistinguishable from 0 and are therefore omitted below.

Greater cue noise results in more reactive choice behavior: The value of cue encoding noise (ε_C) was negatively correlated with the proactive choice index (*r*=–0.57). The greater the cue encoding noise, the more frequently the model infers that it saw the A cue when it in fact saw the B cue (since A is more frequent than B, the Bayesian prior on A is stronger) and the more it makes errors on BX relative to AY trials (the typically reactive mistake).

Greater probe drift rate results in more reactive choice behavior: The drift rate on the probe ($\tilde{a}$) was negatively correlated with the proactive choice index (*r*=–0.69). The higher the signal-

to-noise ratio on the probe relative to the cue, the more accurately the model distinguishes Y from X probes, and the less it is impacted by the left-bias on A trials (caused by the A starting point being closer to the left boundary), reducing AY errors and thereby the proactive choice index.

Smaller difference between A and B response priors results in more reactive choice behavior: The difference between the starting point to threshold ratio for A and B cues ($\frac{{x_0}^B}{z} - \frac{{x_0}^A}{z}$) was negatively correlated with the proactive choice index (*r*=–0.28). The higher the B starting point relative to the A starting point, the more similar its degree of left bias, and the more likely the model is to respond left instead of right on BX trials, making the typically reactive mistake.

To summarize, two main factors can cause the DDP (proactive) model to generate reactive error rates (AY < BX errors): greater signal-to-noise ratio on the probe relative to the cue (due to higher probe drift or higher cue noise) and merging of response priors for the two cues (smaller differences between $\frac{{x_0}^B}{z}$ and $\frac{{x_0}^A}{z}$).

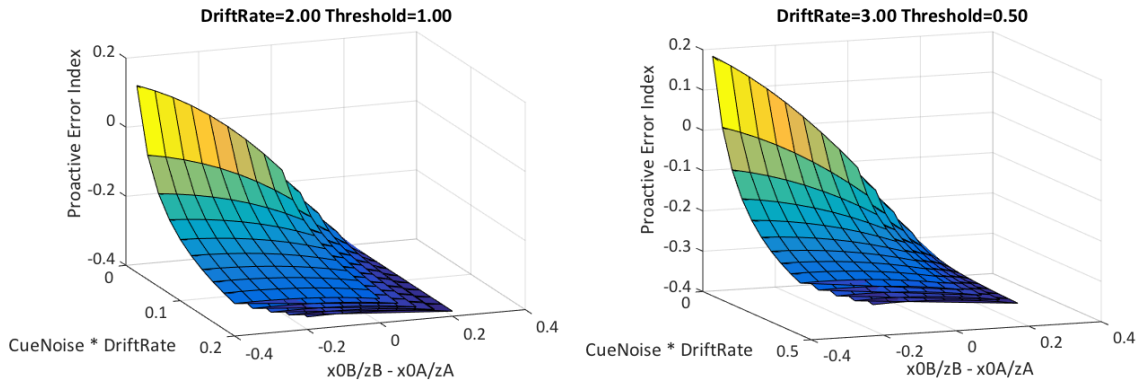


Figure 41 In the DDP model, reactive choice behavior can be induced by greater cue noise, greater drift rate, and smaller differences in response biases between cues. In accordance with previous literature, the proactive choice index was defined as $(AY - BX) \,/\, (AY + BX)$ error rates. This index, plotted on the z-axis, decreases monotonically as a function of two factors: 1) the product of cue noise and drift rate on the probe, and 2) the difference between B and A starting points. Importantly, the index adopted negative values for higher values of both of these factors, suggesting that the proactive DDP model can produce reactive choice patterns.

The purely proactive DDP model cannot generate reactive RT behavior

Unlike the proactive choice index, the proactive RT index decreased only slightly (though significantly) with cue noise (r=–0.11) and probe drift rate (r=–0.08). In contrast, it exhibited a very strong relationship with the difference between A and B response priors (r=–0.92). The smaller the difference in starting points between the two cues ($\frac{{x_0}^B}{z} - \frac{{x_0}^A}{z}$), the smaller (less proactive) the index became.

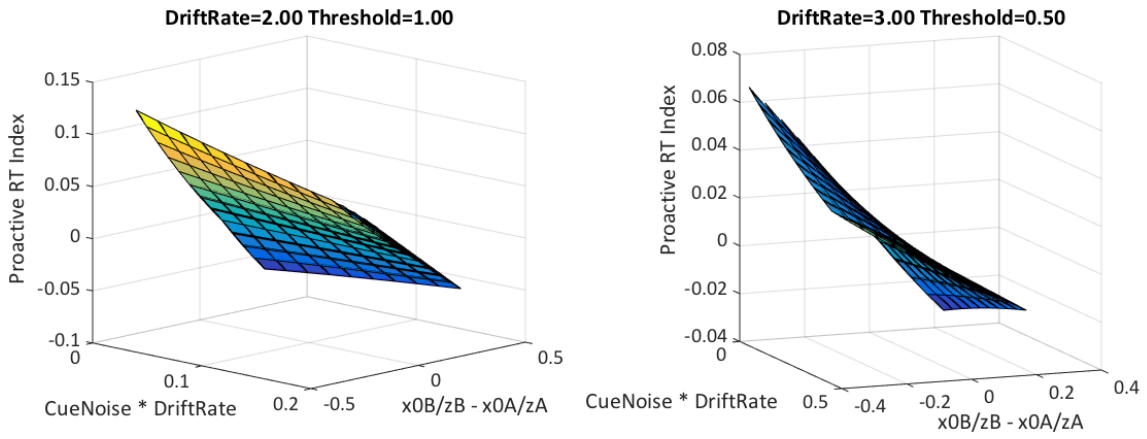


Figure 42 In the DDP model, reactive reaction time behavior is only produced by unrealistic settings of starting point parameters. Unlike the choice index, for which most of the variance was explained by the amount of cue noise and drift rate, most of the variance in the RT index is explained by the difference between B and A starting points. Moreover, the RT index only dropped below 0 (reactive values) when B starting points were higher than A starting points, which is unrealistic given the trial frequencies.

Moreover, unlike the choice index, which mostly adopted negative (reactive) values ($t(7199) = -79.9$, $p \sim 0.00$), the RT index mostly adopted positive (proactive) values ($t(7199) = 54.8$, $p \sim 0.00$), suggesting that the proactive model produces proactive RT indices, in accordance with theoretical predictions. There was only one set of parameter values for which the RT index dipped below 0: when the B prior was more positive than the A prior (when $\frac{{x_0}^B}{z} - \frac{{x_0}^A}{z} > 0$), in other words, when B cues induced a greater left bias than A cues. Such parameters settings are not realistic given the experimental design used in the literature (in all used designs, left responses are far more frequent than right responses after A, whereas left responses are never more frequent than right responses after B). Thus, for all plausible settings of model parameters, the proactive DDP model produces proactive RT indices.

Behavioral Results

As **Figure 43** shows, the proactive-bias manipulation generally succeeded in eliciting more AY than BX errors ($t(119)= 3.35$, $p=0.001$), a pattern conventionally interpreted as evidence of proactive control. The reactive-bias manipulation produced more variable behavior, less consistently eliciting the predicted pattern of more BX than AY errors ($t(119)=1.89$, $p=0.06$). While mean response times varied by trial type (AY errors were faster than AY corrects, and AX errors were slower than AX corrects), they did not vary by condition, possibly because the incentive to respond faster in the proactive condition reduced the overall variability in decision times.

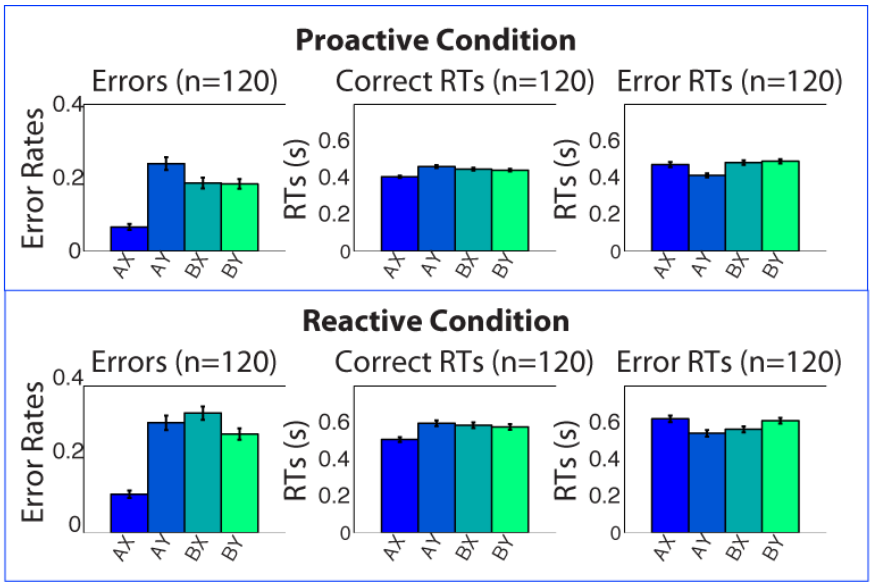


Figure 43 Mean behavior in the two experimental conditions. The top panel shows average behavior in the "Proactive Condition", referring to all blocks in which participants were rewarded for fast and accurate responding, to encourage the use of proactive control. The bottom panel shows average behavior in the "Reactive Condition", referring to all blocks in which participants performed a distractor task during the delay period between cue and probe, to encourage the use of reactive control. Error bars reflect standard errors of the mean across participants. Reaction times on correct and error trials were first averaged within each subject and the average values were then averaged across participants.

Model Goodness of Fit by Experimental Condition and Model Type

The mean pseudo R^2 values ranged between 0.68 for DDP (SD=0.26) and 0.69 for DDC (SD=0.28) (**Figure 44** A). Since the saturated model had four times as many parameters as the theoretical models, the latter received a consistently lower BIC than the saturated model (all $p < 10^{-10}$), suggesting that the DDP and DDC models are a parsimonious description of the data. We discuss more detailed results next.

DDC model fits proactive behavior better whereas DDP model fits reactive behavior better

Considering only the proactive condition, the DDC model fit behavior significantly better than DDP (t(119)=4.1, $p < 10^{-4}$), whereas there was no difference between models for the reactive condition (t(119)=0.07, p=0.94). However, since there were large individual differences in how well the reactive manipulation elicited the desired behavior, we also split participants into two groups based on the sign of their "proactive error index" (AY errors – BX errors). This revealed that proactive behavior (positive index) was significantly better fit by the DDC than the DDP model (t(110)=10.00, $p < 10^{-16}$) whereas reactive behavior (negative index) was significantly better fit by the DDP than the DDC model (t(106)= 3.64, $p < 10^{-3}$, see **Figure 44** B). This pattern was the opposite of what was predicted based on existing interpretations of the proactive error index (e.g., Braver et al., 2009; Gonthier et al., 2016). However, it can be explained by the fact that both models can in fact produce both types of behavior. Given that the DDC model represents cue noise through a low drift rate, it would need broad reaction time

distributions in order to produce BX errors. This was not the case in our data, where participants were encouraged to react quickly. Therefore, the DDP model was able to better capture BX errors through its bit-flip cue noise parameter (which does not affect reaction times.) Similarly, the DDC model's probe noise parameter could better account for proactive AY errors. This effect is accentuated by the lack of reaction time differences across conditions in our data (AY trials were slower than BX in both conditions). In such cases, the encoding noise parameter is better suited for producing error changes without large corresponding RT differences. Future experiments that do not impose a RT deadline are needed to see how model fits change with more variability in the RT distributions.

Realistically, participants may not adopt purely cue-based or probe-based strategies, and evidence accumulation might occur on both stimuli simultaneously at the time of response. Future studies could investigate whether our results generalize to models that capture this interaction (e.g, Shvartsman, Vaibhav, & Cohen, 2015).

Behavior in proactive condition was better fit by all models than behavior in reactive condition

Both models received higher pseudo R^2 scores for the proactive condition than the reactive condition (t(119)=2.09, p=0.04 for DDP and t(119)= 2.73, p=0.007 for DDC, see **Figure 44** A). Given that we did not reward speed or accuracy in the reactive condition, participants were likely more unstable in their strategy across trials. Such behavior is better fit by the saturated model, since it fits a weighted sum of two arbitrary DDMs. However, splitting behavior by the proactive error index did not reveal lower pseudo R^2 scores for reactive behavior (t(216)=0.46,

p=0.64 for DDC), and the DDP model trended towards fitting reactive behavior better than proactive behavior (*t*(216)=1.96, *p*=0.05), suggesting that the poorer fit was not related to higher BX errors, but rather specific to the reactive manipulation. Thus, while the reactive *manipulation* may have promoted more noisy behavior, the models can fit reactive profiles of behavior as well as proactive ones.

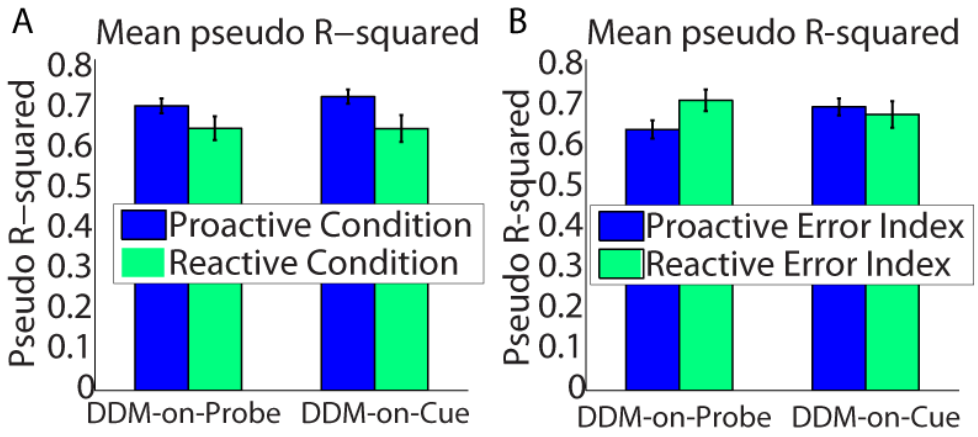


Figure 44 Quality of model fits by condition and proactive index. (A) Model fits revealed that behavior proactive-bias condition was better fit by both DDP and DDC models than behavior in the reactive-bias condition. This may be explained by the fact that participants were better incentivized in the proactive condition, and thus more consistent in their strategy across trials. **(B)** Splitting participants into those with greater AY than BX error rates ("proactive" error index) and those with greater BX than AY error rates ("reactive" error index) revealed that proactive behavioral profiles were better fit by the DDC model, which uses the probe to set parameters for diffusion on the cue, whereas reactive behavioral profiles were better fit by the DDP model, which uses the cue to set parameters for diffusion on the probe.

Landscaping analysis points to more sensitive experimental manipulations for model identification

In addition to the considerations above, the results of the landscaping analysis suggest that the parameters used for experimental design may not have been ones that were maximally sensitive to differences in predictions made by the two model variants.

Comparing the log likelihoods of the two models (on data generated by one of them) for various parameter combinations revealed that they make the most distinct predictions when both encoding noise and drift rate are high, or when both encoding noise and drift rate are low. These are all regimes in which one stimulus (cue or probe) is encoded very noisily, while the other is encoded with high fidelity. To test this, experimenters could parametrically degrade the cue and probe stimuli independently of each other. The landscaping analysis predicts that, in regimes with less peaky reaction time distributions (low drift rate), the DDP model will fit better when the probe uncertainty exceeds the cue uncertainty, and vice versa. In addition, future experiments could test how the proactive error and RT indices (AY – BX) relate to the time spent integrating probe or cue evidence, respectively.

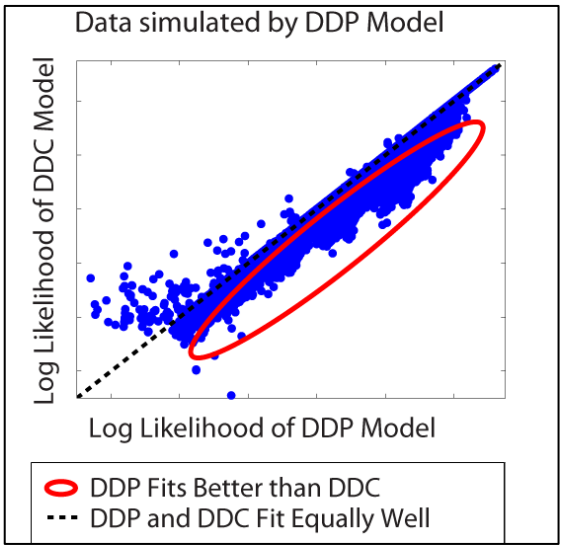


Figure 45 Landscaping Analysis. Simulating behavior from the DDP model and fitting the simulated behavior with both model variants reveals that when the drift rate on the probe is high and encoding noise on the cue is high, the DDP model fits its own simulated behavior better than the DDC model. In other words, when the probe is perceived with *high* signal-to-noise ratio, whereas the cue is remembered with *low* signal to noise ratio, the DDP model generates behavior that the DDC model cannot fit.

Conclusion

Taken together, model simulations, model fits to behavioral data and the landscaping analysis all suggest that reaction times may be a more sensitive measure of control strategy than error rates. First, simulating behavior from the DDP model – a model that perfectly processes the cue in advance of the probe – showed that it could generate reactive choice behavior (AY > BX error rates) but could not generate reactive RT behavior (AY > BX correct RTs) for realistic settings of model parameters.

Second, fitting both model variants to behavioral data revealed that the perfectly proactive model fit reactive choice behavior better, whereas the perfectly reactive model fit proactive choice behavior better. This was the opposite of what was predicted based on previous literature. Further analyses showed that this resulted from the fact that the experimental manipulations only succeeded in affecting participants' choice index, but not their RT index. Based on the simulations above, it is possible that the reactive-bias manipulation did not succeed in encouraging reactive control strategies, given that correct BX RTs never exceeded correct AY RTs on average.

Our results suggest that isolating participants with a proactive vs. reactive reaction time, rather than error rate, index may enable greater separation based on control strategy and facilitate investigations of the neural mechanisms underlying each strategy.

Acknowledgements

This project was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Prior Presentations, Publications and Co-Authorship

This work was completed in collaboration with Robert C. Wilson, Michael Shvartsman and Jonathan D. Cohen. Initial stages of the project have been presented as posters at two conferences:

Lositsky, O., Wilson, R.C., White, J.M., Cohen, J.D. (2013, October). Bayesian model of proactive and reactive control in the AX-CPT. *Computational Psychiatry Conference*, Miami, FL. [***Poster***]

Lositsky, O., Wilson, R.C., Shvartsman, M., Cohen, J.D. (2015, June). A drift diffusion model of proactive and reactive control in a context-dependent two-alternative forced choice task. In Proceedings of the *Reinforcement Learning and Decision Making Conference*, Edmonton, Alberta. [***Short paper and poster***]

References

Bellani, M., & Brambilla, P. (2008). The use and meaning of the continuous performance test in schizophrenia. *Epidemiologia E Psichiatria Sociale*, *17*(3), 188–191. doi:10.1017/S1121189X00001275

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–65. doi:10.1037/0033-295X.113.4.700

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. doi:10.1016/j.tics.2011.12.010

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. *Variation in Working Memory*. doi:10.3233/978-1-60750-044-5-250

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7351–6. doi:10.1073/pnas.0808187106

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459. doi:10.1037/0033-295X.100.3.432

De Pisapia, N., & Braver, T. S. (2006). A model of dual control mechanisms through anterior cingulate and prefrontal cortex interactions. *Neurocomputing*, *69*(10–12), 1322–1326. doi:10.1016/j.neucom.2005.12.100

Fernandes, H. L., Stevenson, I. H., Phillips, A. N., Segraves, M. A., & Kording, K. P. (2014). Saliency and saccade encoding in the frontal eye field during natural scene search. *Cerebral Cortex*, *24*(12), 3232–3245. doi:10.1093/cercor/bht179

Gonthier, C., Macnamara, B. N., Chow, M., Conway, A. R. A., & Braver, T. S. (2016). Inducing proactive control shifts in the AX-CPT. *Frontiers in Psychology*, *7*(NOV). doi:10.3389/fpsyg.2016.01822

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. doi:10.1146/annurev.neuro.24.1.167

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230. doi:10.1016/j.jmp.2009.02.003

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and

the informativeness of data. *Cognitive Psychology*. doi:10.1016/j.cogpsych.2003.11.001

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi:10.1037/0033-295X.85.2.59

Riccio, C. A., Reynolds, C. R., Lowe, P., & Moore, J. J. (2002). The continuous performance test: A window on the neural substrates for attention? *Archives of Clinical Neuropsychology*. doi:10.1016/S0887-6177(01)00111-1

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, *20*(5), 343–350. doi:10.1037/h0043220

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Shvartsman, M., Vaibhav, S., & Cohen, J. D. (2015). A Theory of Decision Making Under Dynamic Context. *Advances in Neural Information Processing Systems 28*, 1–9. doi:10.1371/journal.pone.0009308

General Discussion

In this thesis, we investigated the role of contextual representations in memory, learning and decision-making. In Study 1, we investigated how contextual changes might act as "event boundaries" that cause memories to cluster within events and separate across events, and how such separation of memories by context influences retrospective time perception. In Study 2, we investigated how people learn whether it is advantageous to separate memories by context or to combine them across contexts. We manipulated the degree to which two contexts made similar predictions about stimuli and responses, and found that people combined predictions across contexts to the extent that they were similar. In Study 3, we used a diffusion model to test predictions of the "dual mechanisms" framework, which posited that in addition to using context to predict upcoming stimuli, people might use stimuli to retroactively retrieve the representation of the context.

Below, we briefly review the motivations and findings of each of the three studies and discuss outstanding questions for future research.

Contributions of Study 1

Three findings from the literature, reviewed above, motivated Study 1. First, multiple studies have shown that changes in the spatial environment, task instructions and type of processing performed on stimuli, and event boundaries (which themselves tend to be a form of contextual change) all caused people's estimates of duration to dilate in retrospect. In other words, if participants were asked to estimate the duration of an experience retrospectively,

their estimates seemed to scale with the amount of context change that had occurred during the experience. Second, studies that measured recognition memory or free recall of stimuli from the experiment have found that greater context changes helped participants to better remember at which time the stimulus had occurred. Moreover, studies have found that event boundaries cause associative memories to weaken across events but strengthen within events, suggesting that contextual boundaries trigger the formation of "episodes" in long-term memory, producing clustering of memories by context. Third, a wealth of neuroscience studies have found that the medial temporal lobes and lateral prefrontal cortex play an important role in encoding and remembering associations between stimuli and their contexts, and found that the temporal order or temporal context of stimuli was better remembered when patterns of activity in these regions (measured using electrophysiology of fMRI) tended to change more. Based on these three findings, we hypothesized that changes in the neural activity patterns in the MTL and lateral PFC might encode changes in contextual information, and that such pattern changes during encoding of a story should correlate with people's estimates for how long different intervals in the story lasted.

The experimental results of Study 1 supported this hypothesis. First, we found that participants' estimates of how much time had passed between two clips from the story were longer when more event boundaries had occurred between the two clips. These findings replicated the results of previous studies, showing that contextual change during an interval lengthened duration estimates for that interval. Second, we found that the degree of neural pattern change in the right anterior temporal lobe between two clips (during initial listening of the story) predicted participants' estimates of how much time passed between those two clips

at the end of the experiment. This finding supported our hypothesis that patterns in brain regions particularly important for encoding and remembering stimulus-context associations changed more between clips for which the temporal distance was later overestimated. Third, we found that both of the effects above survived a "control" analysis in which the effect of perceptual similarity was regressed out of participants' duration estimates. The control analysis suggested that both duration estimates and pattern changes in the anterior temporal lobe are influenced by changes in more abstract factors than the immediate similarity between the two clips, factors that are held in memory while the clips are being heard for the first time.

Since we did not track neural activity during the time perception and memory test, our results do not speak directly to which strategy participants used to estimate the durations of intervals they experienced many minutes ago. It is possible that participants might only retrieve contextual memories from the two clips (from the beginning and end of each interval). However, it is also possible that they could retrieve memories from the entire interval, replaying what happened between the two clips. Future studies could test this by measuring whether patterns of activity during the retrieval phase, when participants are estimating durations, show reinstatement of patterns from the interval between the two story clips.

Despite this limitation, Study 1 contributes to the existing literature in four important ways. First, it builds on a study by Ezzyat & Davachi (2014) that used neuroimaging to investigate how separation of memories by contextual boundaries affects retrospective time perception. The paradigm used in that study was not entirely retrospective (in Ezzyat & Davachi (2014), participants became aware of the need to estimate time after the first block) and the intervals probed did not exceed one minute. In contrast, our study used a completely retrospective

paradigm, in which participants encoded the entire story before being asked to estimate time, and all interval durations were over a minute long. Second, as we far as we know, it was the first neuroscience experiment to investigate how time is perceived in naturalistic paradigms (a radio story). In all previous paradigms, contextual information could be clearly separated from the stimuli on which the task was performed and could constitute the background color, the category of the stimulus or the type of task performed. In contrast, the naturalistic story presents continuity between stimuli and contexts, which is more representative of real life. Since then, an additional neuroimaging study has used naturalistic video stimuli to study prospective time estimation (Cohen, Henin, & Parra, 2017). Third, the experimental design allowed us to obtain multiple retrospective duration estimates within the same participant, without revealing the need to measure time while the interval was experienced (as far as we know, only a handful of studies had leveraged similar designs previously; see Faber & Gennari, 2015; Noulhiane, Pouthas, Hasboun, Baulac, & Samson, 2007). Finally, we were able to track changes in patterns of neural activity on the scale of minutes thanks to a methodological advance: we developed an fMRI data preprocessing pipeline that allowed us to remove the maximal amount of low-frequency scanner and physiological noise, while preserving the maximal amount of neural signal that changed on slow time scales.

How findings from Study 1 inform Study 2: Do duration estimates increase when people "split" representations by context?

The results of Study 1 broadly support propositions in the literature that retrospective duration estimates could be used as an index of contextual change. However, in that study,

contextual change involved changes in the external environment (spatial location, characters present, topic of conversation) as well as internally maintained situation models. An important direction for future research would be to verify whether duration estimates track separation of task representations as participants learn to cluster task representations by context. Specifically, in Study 2, we tested whether people "merge" predictions across contexts when two contexts make sufficiently similar predictions. We found that representations appeared clustered by context only if the contexts predicted sufficiently different stimulus and response statistics. Future studies could establish a link between Study 1 and Study 2 by testing how separating priors by context (vs. merging priors across contexts) influences retrospective time estimates. For instance, would participants who learned different priors for A and B contexts believe the experiment lasted longer compared with participants who learned general response priors? (Presumably, participants who are splitting representations by context are experiencing greater "context change".) Analogously, in the Gershman and Niv (2013) study, would participants who learned to use color information to predict the number of circles believe the experiment lasted longer than participants who learned to combine predictions across colors? In such investigations, it will be important to control for changes in the actual stimuli presented and the amount of prediction error across trials, since conditions where the contextual cues predict different stimulus distributions also tend to have greater perceptual change (duration estimates could be lengthened simply because the stimuli themselves change more). A well-controlled experiment could use a similar design to Experiment 2 in Study 2, in which the stimulus statistics change mid-way through the experiment, such that both groups of participants (those who learned to "merge" and those who learned to "split") receive exactly

the same stimuli, and any change in their behavior must arise from statistics they learned earlier in the experiment. A convincing finding would be if participants who learned to "split" representations by context would display greater duration estimates than participants who learned to "merge", even in the second part of the experiment, where the environment statistics are exactly identical.

Another paradigm that could be used to investigate such questions is the "Whack a Mole" experiment used in Qian, Jaeger, & Aslin (2016). In their study, participants attempted to "whack a mole" that could appear in one of four locations. Unbeknownst to the participants, the probability distribution of the mole over the four locations was stable across trials within a "bundle" (an event), and changed unexpectedly across bundles. There were four bundles used in total, such that previously encountered bundles could recur later in the experiment. Qian et al. (2016) provide evidence that participants' behavior adapted to these unequal probabilities over the four locations, adapted to the changes in these probabilities across bundles, and showed improved performance when a bundle recurred a second time, suggesting that participants retained memory of each bundle. However, importantly, participants did not seem aware of this learning (at debriefing, they were unaware that there were different bundles and that bundle identity could change suddenly). It would be fascinating to explore whether duration estimates might be longer in such paradigms, where participants' behavior shows evidence that multiple "categories" (bundles) were learned, despite the fact that participants are entirely unaware that the experiment is segmented into bundles.

To conclude, additional research is needed to verify which types of contextual changes influence retrospective duration estimates. If duration estimates lengthen in conditions that

bias participants towards learning two categories, rather than a single category, then such estimates could constitute an important tool for research on structure learning.

Contributions of Study 3

In Study 3, we used the same drift diffusion model of context-dependent behavior used in Study 2 to address a different question about how within-trial dynamics of context processing influence behavior. Previous studies (Reynolds, Zacks, & Braver, 2007) had proposed that there might be two ways in which context-dependent behavior can arise in tasks like the AX-CPT: the contextual cue could be processed in advance and used to guide responding to the probe (proactive strategy), or the contextual cue might be forgotten, but the probe itself could be used to guide retrieval of the contextual cue from longer-term memory (reactive strategy). Despite the large body of work providing evidence for the dual mechanisms framework (Gonthier, Braver, & Bugg, 2016; see Braver (2012) for a review), Study 3 was the first to build a computational model of proactive and reactive strategies in the AX-CPT task, and the first to use this model to verify how these control strategies affect behavior.

A multitude of studies has used an index combining error rates and mean correct reaction times on AY vs. BX trials in the AX-CPT as a measure of proactiveness (Gonthier, Macnamara, Chow, Conway, & Braver, 2016). The worse people perform on AY relative to BX trials, the more proactive they are considered to be, and vice versa, the worse people perform on BX relative to AY trials, the more reactive they are considered to be. To verify how the behavioral index is impacted by various mechanisms, we built a proactive and reactive variant of the drift diffusion

model. Both model variants combine cue and probe information to make a correct response, but differ in which of the two (the cue or the probe) is processed first. We operationalized proactive control as a model that always processes the cue entirely before the probe, and uses the cue to set parameters for the decision on the probe. Reactive control was operationalized in the exact opposite way. In the proactive variant, the cue was completely processed by the time the probe arrived, such that the drift rate (evidence accumulation towards the correct decision) was supposed to reflect only the processing of the probe. Simulations from this proactive variant show that it can generate reactive error rates (BX > AY errors) using entirely realistic settings of two parameter sets: 1) higher cue noise and higher drift rate on the probe, and 2) overly similar starting points for A and B cues. Regarding the first parameter set, the fact that increasing cue noise increases the degree to which error rates look "reactive" presents an important concern, as it suggests that a participant with noisier representation of the cue (potentially due to lower working memory capacity) could look reactive even when using a perfectly proactive strategy. By perfectly proactive strategy, we mean that the cue is still being processed entirely in advance of the probe, only it is being processed with lower accuracy and fidelity, such that it appears as a weaker influence on the decision. The dual mechanisms of control (DMC) framework does suggest that participants with lower working memory capacity would be more likely to use reactive strategies. However, our model shows that no additional reactive strategy is needed to account for higher BX than AY mistakes – these could arise from a purely proactive one with low working memory capacity. As for the second parameter set, as we discuss at length in the next sub-section, similar starting points for A and B cues could arise from merging response probabilities across cues (as shown in Study 2), rather than from a

reactive strategy. Thus, Study 3 shows that BX > AY *error rates* do not necessarily reflect reactive control.

In contrast, the reaction time (RT) part of the proactive-reactive index was more robust. A proactive model, which always processed the cue in advance of the probe and used the cue to set the parameters for the decision on the probe, was unable to produce a RT index that looked reactive (where correct BX trials are slower on average than correct AY trials). A reactive RT index could only be produced by forcing the B Prior to be higher (more left-biased) than the A Prior, which is unrealistic given the trial frequencies. These simulations suggest that slower BX than AY trials reflect later processing of the cue relative to the probe, which we believe is the essence of a reactive strategy. If this prediction of our model is correct, then the neural mechanisms of proactive and reactive control might be better investigated if researchers employ the mean BX – mean AY RTs as an index of strategy.

In future work, we plan to test this prediction by obtaining a large amount of neural and behavioral data from laboratories that have investigated mechanisms of proactive and reactive control using the AX-CPT. Behaviorally, we can test whether using the reaction time portion of the index strengthens the correlations between the behavioral index and other psychological measures that have been correlated with this index, such as tests of working memory capacity, responses on psychological questionnaires testing trait anxiety, and many others. Using neuroimaging data collected by other labs, we can also test whether using the RT portion of the index produces stronger correlations with time courses of activation in regions that have been found to distinguish proactive and reactive strategies (e.g., regions of the lateral PFC and dorsal ACC).

How findings from Study 2 inform Study 3

By characterizing the effects of trial frequencies on AX-CPT behavior and model fits, Study 2 sheds light on potential confounds in the behavioral index of proactive control. In Study 2, we showed that manipulating the degree to which A and B make similar predictions about the response dimension changes the degree to which participants group response probabilities across cues. Grouping response probabilities across cues manifested as more similar starting point parameters for A and B cues (starting point parameter estimates were obtained from fitting our drift diffusion model to behavior). Importantly, in Study 3, simulations from the same drift diffusion model showed that varying the similarity of A and B starting points shifts the error rate index that researchers have used to measure the degree of proactive vs. reactive control.

The fact that closer starting points across cues produce more reactive choice behavior is an important concern for interpreting both across-participant and within-participant variability in the error rate index. It suggests that participants who make more BX than AY mistakes may do so because they are merging response probabilities across cues more, rather than using a more reactive strategy. Of course, under some interpretations of the dual mechanisms framework, it could be argued that merging probabilities across cues is a form of reactive control in itself, because it is a strategy in which context is used less to guide predictions. However, we argue that the dual mechanisms framework addresses only within-trial dynamics, whereas Study 2 showed that merging of starting points across cues could result from probability learning across many trials. Therefore, so long as researchers use the error rate component of the proactive

index (AY-BX errors) as a measure of within-trial dynamics, the results of Study 3 show that the index is confounded with across-trial probability learning.

Conclusion

To conclude, the above studies have investigated the mechanisms by which context influences behavior across many time frames. Study 1 showed that context can influence the way in which memories are clustered, which affects estimates of elapsed time many minutes later. Study 2 showed how participants learn to either merge or separate predictions by context incrementally over many trials, and how such learning influences both task performance and explicit probability estimates at the end of the experiment. Study 3 modeled different dynamics of context processing, either in advance of the stimulus or retroactively, and attempted to discern such dynamics from behavior.

In addition to deepening our understanding of the above topics, these studies show the value of combining various approaches. Study 1 combined insights from experimental psychology studies of prospective and retrospective timing with neuroscience findings about the role of the MTL and PFC in representing mental context. It sought to fill the gap in the neuroscience literature on time perception, which had only investigated intervals that were a few seconds long and had only used prospective paradigms (in which time had to be measured deliberately). Study 1 also employed neuroimaging analysis methods developed for leveraging the highly reliable signal time courses induced by naturalistic stimuli and adapted them for investigating neural activity on longer time scales with fMRI. Study 2 leveraged insights from perceptual studies of category learning and cue combination and tested whether similar principles of Bayesian inference can explain how predictions are separated by context and used in context-dependent decisions. Using a drift diffusion model to fit and predict behavior was

instrumental to the success of this study; it not only allowed far more powerful analyses of the behavioral data (the starting point parameter in the model revealed much clearer patterns than either reaction time or accuracy data on their own, since the model fits took into account the entire distribution of reaction times) but also provided us with predictions for how an optimal agent would set its priors differently by context. These model predictions enabled us to discover that participants' decision starting points were in fact optimal given the probabilities they had learned, suggesting that merging across contexts arises from limitations in learning, rather than limitations in implementation given learned priors. Finally, the same diffusion model was instrumental in Study 3, where it enabled us to test whether proactive and reactive strategies produced the behavioral patterns hypothesized by the literature. Using insights from the model, we may be able to distinguish influences of working memory failure, across-trial probability learning and context processing dynamics on behavior.

References

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. doi:10.1016/j.tics.2011.12.010

Cohen, S. S., Henin, S., & Parra, L. C. (2017). Engaging narratives evoke similar neural activity and lead to similar time perception. *bioRxiv*, 1–16. doi:10.1101/104778

Ezzyat, Y., & Davachi, L. (2014). Similarity breeds proximity: Pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron*, *81*, 1179–1189. doi:10.1016/j.neuron.2014.01.042

Faber, M., & Gennari, S. P. (2015). In search of lost time: Reconstructing the unfolding of events from memory. *Cognition*, *143*, 193–202. doi:10.1016/j.cognition.2015.06.014

Gonthier, C., Braver, T. S., & Bugg, J. M. (2016). Dissociating proactive and reactive control in the Stroop task. *Memory & Cognition*, (February), 778–788. doi:10.3758/s13421-016-0591-1

Gonthier, C., Macnamara, B. N., Chow, M., Conway, A. R. A., & Braver, T. S. (2016). Inducing proactive control shifts in the AX-CPT. *Frontiers in Psychology*, *7*(NOV). doi:10.3389/fpsyg.2016.01822

Noulhiane, M., Pouthas, V., Hasboun, D., Baulac, M., & Samson, S. (2007). Role of the medial temporal lobe in time estimation in the range of minutes. *Neuroreport*, *18*(10), 1035–8. doi:10.1097/WNR.0b013e3281668be1

Qian, T., Jaeger, F. T., & Aslin, R. N. (2016). Incremental implicit learning of bundles of statistical patterns.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*(4), 613–43. doi:10.1080/15326900701399913