TRACKING THE SUMMARY STATISTICS IN LONG-TERM MEMORY

Haiyun Zeng

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Dr. Sharon L. Thompson-Schill

Christopher H. Browne Distinguished Professor of Psychology

Graduate Group Chairperson

Dr. Russell Epstein

Professor of Psychology

Dissertation Committee

Dr. Sharon L. Thompson-Schill, Christopher H. Browne Distinguished Professor of Psychology

Dr. John C. Trueswell (Chair), Professor of Psychology

Dr. Anna C. Schapiro, Assistant Professor of Psychology

TRACKING THE SUMMARY STATISTICS IN LONG-TERM MEMORY

Dedicated to H. Zeng and Y. Li.

ABSTRACT

TRACKING THE SUMMARY STATISTICS IN LONG-TERM MEMORY

Haiyun Zeng

Sharon L. Thompson-Schill

Decades of research have demonstrated humans' extraordinary ability to extract summary statistics across individual experiences. Less is known about how exactly the items contribute to the summary statistics and how the relationship between memory for the items and memory for the summary statistics evolves and changes over time. I propose that memory of summary statistics that are initially extracted from individual instances starts to guide memory for individual items over time, and not all items contribute to the summary statistics equally. Sources of item distinctiveness influence the summary statistics extraction in terms of the contribution of each item and the accuracy of summary statistics. The three empirical chapters enlighten our understanding of summary statistics extraction in long-term memory by bridging fields ranging from perception and memory to emotion and motivation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1: General Introduction

What comes to your mind when you think of an average bear, and how does this representation relate to the individual bears that you have encountered? The "average bear" that first comes to your mind may be similar to the bears you know, such as your teddy bear or the bear you saw on TV yesterday. However, it is not exactly the same with all of them. How does this memory of the average bear form from the individual bears you have seen?

How learners extract summary statistics from individual instances is a fundamental question in psychology. Research in various fields in cognition, such as perception, memory, and concept, has studied summary statistics as the "gist", "prototype," and "schema" of individual experiences (Armstrong et al., 1983; Ghosh & Gilboa, 2014; Lewis & Durrant, 2011; Rosch & Mervis, 1975). Among the summary statistics, "average" plays a special role in human cognition and behavior. This is demonstrated by extensive work in various fields of psychology, such as concepts, perception, and attractiveness. For example, Memory for instances of pre-existing semantic categories (e.g., strawberry) is biased towards the center of the category (e.g., average size of fruit) (Hemmer & Steyvers, 2009). Perception work shows that people's memory of newly learned items such as the size of circles and emotional expressions is biased towards the average of the group of these instances (Brady & Alvarez, 2011; Corbin & Crawford). Average faces are perceived to be more attractive (Valentine et al., 2004), and this preference for average may be linked to their visual experience with these faces within cultures (Apicella, et al., 2007).

Given the importance of average, less is known about how such average is formed and computed from observing, experiencing, and learning about individual items and events over time. In particular, how does each individual experiences contribute to the summary statistics in long-term memory? If humans can extract summary statistics from individual items, will certain individual items contribute to the summary statistics more than the others? What factors about the items influence such computation?

Perception literature has offered intriguing theories, paradigms, and models for understanding how much each item contributes to the summary statistics, and the factors that influence this computation. Research in ensemble perception operationalizes summary statistics as a memory of the "average" across items that vary on a continuum, for instance, the average size of many circles that vary in size. Based on the paradigms, ensemble perception work has developed models to estimate the weights of the individual items and has found evidence that not all items contribute to the summary statistics equally (Alvarez, 2011; Whitney & Leib, 2018). Factors such as displayed order, accuracy, deviancy, and attention, will influence the weight of items in the summary statistics. Items that are presented at the beginning and at the end of the sequence of the items weigh more in the summary statistics (Hubert-Wallander & Boynton, 2015; Tong et al., 2019). Items with more reliability and more similarity to the summary statistics weigh more in the summary statistics (de Gardelle & Summerfield, 2011). Items that received more attention shift the summary statistics of a group of emotional faces more (Ying, 2022). Moreover, "outlier" items, which are items that are more deviant from the other items, are discarded in forming the summary statistics (Haberman & Whitney,

2010). These results suggest that not all items weigh equally in forming the summary statistics and the properties of the items influence the weights.

In long-term memory research, there have been some studies that suggest that the property of the items may change their contribution to the summary statistics. For example, items that are "outliers" relative to the spatial pattern of all other items in learning can greatly disrupt rodents' pattern identification process (Richards et al., 2014). Humans' memory of summary statistics changes after learning an item that is inconsistent with other items and also this change is sensitive to consolidation time (Richter et al., 2019). However, not much work has systematically investigated how exactly these outliers contribute to the summary statistics, similar to research in ensemble perception.

Exploring the differential weighing of items in long-term memory can enlighten our understanding of the possible common mechanisms of summary statistics computation between perception and cognition. These two fields study summary statistics in parallel and there are already striking similarities in the findings. In both fields, learners can preserve accurate memory of summary statistics while losing the accuracy of item memory. In long-term memory literature, it was discovered that humans were capable of reporting the prototype even when they forgot the individual items one year after learning the individual items (Lutz et al., 2017). On the other hand, Perception researchers found that humans can rapidly compute the summary statistics from seeing a group of individual instances. For example, participants can accurately report the average circle size after briefly (e.g., 500 ms) seeing a dozen of circles of varying sizes (Ariely, 2001). Despite the similarities, it is unclear whether these findings in perception and memory are governed by the same mechanisms. Currently, perception literature has developed

systematic methods for understanding the weights of the items in the summary statistics. Investigating how the items contribute to the summary statistics using similar paradigms in long-term memory will offer new evidence for the connection between the two fields.

Furthermore, exploring this mechanism of differential weighing in long-term memory could be important for other fields in psychology, such as categorization. According to the prototype model in categorization literature, people extract the "central representation" and may utilize this representation to categorize and generalize new exemplars (Homa et al., 2019; Posner & Keele, 1968; Smith & Minda, 1998). When testing the theory that learners used prototypes to make categorization decisions, categorization research does not usually examine how the exemplars contributed to the prototype. However, recent work in this field suggested the way the exemplars are presented (i.e., repeated exposure) may make a difference in learners' categorization performance (Homa et al., 2019). Thus, understanding how this prototype is computed from individual items may improve the performance of the prototype model. For example, the prototype model in categorization literature often assumes equal contribution from the items to the summary statistics. The prototype defined in these studies usually assumes equal weights over the items (Nosofsky, 1987; Smith & Minda, 2000; Tong et al., 2019). If similar to ensemble perception literature, when the items were presented influenced their contribution to the summary statistics, including this assumption of differential weights may improve the prototype models and thus add to the categorization literature.

Investigating the mechanism of differential weights in long-term memory can provide insights into the adaptive values of summary statistics extraction. If some items are

remembered more accurately compared to other items, then giving more weight to these items may improve the accuracy of the overall summary statistics (Alvarez, 2011). However, differential weights in items can distort the summary statistics and lead to biases. If an outlier item greatly disrupts rodents' seeking pattern for platforms (Richards et al., 2014), wouldn't this lead to a lower survival chance because the majority of the resources still follow the overall pattern? Providing evidence of what factors will lead to differential weights will be the first step toward understanding the adaptive values of summary statistics extraction.

Chapter 2 aims to understand the role of the formation of memory of summary statistics in long-term memory and how the relationship between the memory of summary statistics and the memory of individual instances evolve and influence each other over one to two months. Prior research in ensemble perception shows that learners discount the outliers when forming the summary statistics (Haberman & Whitney, 2010), whereas scarce evidence from long-term memory work shows that these items greatly distort the summary statistics more than other items (Richards et al., 2014). To resolve the discrepancy, we examined the influence of an"outlier" item on the summary statistics in long-term memory. Inspired by perception literature, this study operationalized items to be spatial locations on a screen and summary statistics to be the center of these locations. Moreover, research in ensemble perception shows that memory of the items is biased toward the summary statistics of the groups of items (Brady & Alvarez, 2011; Corbin & Crawford 2018). However, not much is known about how the occurrence of an outlier will influence this bias. Our study tracked the influence of an outlier item on the memory of the summary statistics over time. We found that the existence of an outlier

changed the memory of summary statistics. Specifically, the outlier consistently weighed more in memory of summary statistics compared to other items over time. However, all the items were increasingly biased towards the center without the outlier over time. These results shed light on how items that are more deviant influence the overall summary statistics in long-term memory.

Chapter 2 provides a start for us to understand the contribution of items to the summary statistics quantitatively. The findings of the outlier lead to more questions: does the outlier weigh more because it is more salient compared to other items? Will other sources of distinctiveness influence how much the items contribute to the gist memory? Will properties of items, such as attention, accuracy, and frequency influence their contribution to the summary contribution? Chapter 2 includes a weighted model on items based on accuracy but it did not show an influence of accuracy on the weights, potentially because the variance in item accuracy occurred naturally and we did not directly manipulate the accuracy. In order to understand how the properties of items influence their contributions to the summary statistics, it will be useful to have a study that systematically manipulates the properties of the items and disentangle their influence. In particular, the reward will be a particularly interesting source of distinctiveness, because work on motivational learning has shown that it recruits a similar brain network as gist extraction (Murty et al., 2016; Tse et al., 2007, 2011; Zeithamova et al. 2012).

Chapter 3 aims to further disentangle the influence of item properties on their contributions to the summary statistics, such as rewards, attention, and frequency. For example, if your teddy bear always gives you a warm emotional reward, will it contribute more to your memory of an average bear? Much evidence has shown that reward during

encoding enhances memory for the items (Murty et al., 2011), but not much research examines its consequence on the summary statistics. Reward, by increasing attention or accuracy of particular items, may increase their weights in summary statistics (Alverez, 2011). Alternatively, reward, by recruiting a similar brain network shown to be related to summary statistics extraction (Tse et al., 2007, 2011; Zeithamova et al. 2012), may facilitate the integration of items and thus preserves the weights of the reward items to match the other items (Clewett & Murty, 2019). On the contrary, the reward may result in a separation between the emphasized item and the other items, which may result in a discard of the emphasized item (Haberman & Whitney, 2010). Our study aims to disentangle these possibilities by measuring the weights of the rewarded items in contributing to the summary statistics, using the same spatial memory paradigm as in Chapter 2. We discovered that the rewarded item, despite a higher accuracy compared to other items with no reward, did not contribute to the summary statistics more. This is in contrast with other conditions which strengthened item memories with other reinforcement, such as frequency and attention. Items with higher frequency contribute to the summary statistics more. Our results are more consistent with the integration account.

Moreover, ensemble perception literature suggests that the mechanism of summary statistics extraction may vary by feature domains, from low-level processing of size and orientation to high-level processing of facial expression (Haberman & Alvarez, 2015). Little work has explicitly tested whether the summary statistics extraction in long-term memory varies across feature domains as well. For example, will the extraction of overall emotions follow a similar rule to the extraction of a pattern of spatial locations? In addition, it is underexplored in long-term memory how the feature domains interact with

each other in forming summary statistics. The broaden-and-build theory proposes positive emotions will broaden the scope of attention and thus encourage global processing, whereas negative emotions narrow the scope of attention and are likely to lead to local processing of emotions (Fredrickson & Levenson, 1998). Ensemble perception work demonstrates a similar difference in positive and negative emotions in the extraction of summary statistics (Peng et al., 2022). However, not much work has been done to explicitly test this theory in the summary extraction of emotions in long-term memory, potentially because of a lack of opportunities to directly measure and compare negative and positive emotions in long and emotionally varied events. Examining whether emotions also influence the extraction of summary statistics in long-term memory will broaden the scope of this theory.

Chapter 4 aims to understand humans extract the summary statistics of emotions from daily experiences over an extended period of time, and whether this extraction varies in positive and negative emotions. Prior literature on emotional memory shows evidence that the "peak" of an experience, which is the most intense and extreme moment, uniquely contributes to the overall evaluation of the emotion, for events in the lab or in real life that lasts a shorter amount of time. Our study intended to integrate this line of work with other long-term memory research in cognition that studies how humans extract summary statistics from individual items. We tracked 160 MTurk participants' daily emotions during the first two months of the COVID-19 pandemic, their recall of these daily emotions, and their recall of their average emotions at a 1-week and 1-month delay. Our analysis showed that the "peak" still uniquely contributes to the summary statistics over an extended period of time for negative emotions and the recall for summary

statistics was more negative than participants' experience, although we did not find similar effects for positive emotions. On the other hand, we found that the recall of date-specific emotions was more negative than the actual date-specific emotion, but this negative bias decreased over time. These findings provide new insights into the extraction of summary statistics in memory and emotion.

Through three empirical chapters, we provide new evidence for summary statistics extraction in human memory in terms of how the properties of items contribute to summary statistics. Our findings bridge research in areas of psychology ranging from perception and memory to emotion and reward, providing important new insights into our ability both to learn about distinct events and to generalize across similar experiences in different domains in psychology.

Chapter 2: Tracking the relation between gist and item memory over the course of long-term memory consolidation

Abstract

Our experiences in the world support memories not only of specific episodes but also of the generalities (the 'gist') across related experiences. It remains unclear how these two types of memories evolve and influence one another over time. In two experiments, 173 human participants encoded spatial locations from a distribution and reported both item memory (specific locations) and gist memory (center for the locations) across one to two months. Experiment 1 demonstrated that after one month, gist memory was preserved relative to item memory, despite a persistent positive correlation between them. Critically, item memories were biased towards the gist over time. Experiment 2 showed that a spatial outlier item changed this relationship and that the extraction of gist is sensitive to the regularities of items. Our results suggest that the gist starts to guide item memories over longer durations as their relative strengths change.

Introduction

Our experiences in the world are perceived and remembered both as individual items, events, and episodes, and also as aggregated collections or sets of related items with common properties. For example, one can remember seeing a brown bear at the zoo, a polar bear at an aquarium, an animated bear in a Winnie the Pooh movie, and on and on; but one also can readily understand the phrase "smarter than your average bear" by aggregating over those individual experiences. A fundamental question in cognitive

science is how we extract summary statistics from individual instances, both during perception and working memory (where aggregated information is often referred to as an "ensemble") and during episodic encoding and retrieval (where aggregated information is often referred to as a "schema" or as the "gist" of an experience)[1]. In addition, researchers have attempted to characterize how memory for these types of information changes over time. For example, studies of long-term memory in both humans and animals have demonstrated that gist memory persists or even improves over time, whereas memory for the individual items from which the gist is built fades (Posner & Keele, 1970; Richards et al., 2014).

What do these observations of temporal dissociations tell us about the relation between item memory and gist memory? On the one hand, a persisting gist memory with less accurate item memory is often taken as evidence that a gist representation becomes independent of individual item representations as it is abstracted during encoding (Posner & Keele, 1970) or through consolidation (Richards et al., 2014). On the other hand, a persisting gist memory with less accurate item memory is not sufficient evidence for the independence of a gist representation: Even when item memories become noisy and less accurate, they still can retain enough information to support a relatively intact memory of gist at retrieval (Alvarez, 2011; Squire, Genzel, Wixted, & Morris, 2015).

Disentangling these two possibilities based on existing evidence is difficult, because previous studies do not have a direct measurement of the gist information retained in item memories. In this study, we developed a paradigm to test item memory, gist

[1] The "gist" in this paper means "generalities across individual instances", instead of "lack of details", "less precise", or "abstract" as in fuzzy trace theory (Brainerd & Reyna, 2002).

memory, and "estimated" gist memory, which is an estimate of gist memory given the assumption that it is assembled from individual memories of constituent items. Ensemble perception research was a source of inspiration in developing such a paradigm. Studies of rapid perception of complex visual arrays reveal precise representations of gist (i.e., ensemble statistics) with less accurate item memory retrieval in working memory (Ariely, 2001). In order to investigate the relation between item and gist, ensemble perception paradigms often operationalize the gist as the average representation across instances.

Following this reasoning, we operationalize item memory as a set of landmarks (e.g., restaurant and university) whose locations are clustered together on a screen, and gist memory as the center for these landmark locations. In our paradigm, participants learn this set of landmarks, and are then asked to report the spatial center of these landmarks and recall the locations of each landmark. Importantly, an "estimated center" can be computed based on their retrieval of individual items, and its accuracy can thus reveal the amount of gist information available in item memories. Thus, we can investigate the relation between gist memories and item memories over the course of long-term memory consolidation by measuring the relationship between the accuracy of the reported center and the estimated center. A positive correlation between estimated and reported center accuracy could mean that participants' gist memory was still supported by individual item memories, or that the gist was influencing the retrieval of items.

To probe the direction of this relationship, we developed a gist-based bias measurement, an approach borrowed from research on hierarchical clustering models and from semantic memory, which both reveal how gist memory influences memory for specific items (Brady & Alvarez, 2011; Hemmer & Steyvers, 2009; Tompary &

Thompson-Schill, 2021). This measure of bias indicates the magnitude of the particular direction in which the items were attracted. Theories suggest that this influence reveals a reconstructive memory retrieval process (Brady, Schacter, & Alvarez, 2015; Hemmer & Steyvers, 2009; Schacter, Guerin, & Jacques, 2011) that depends on the relative strength of item and gist memory (Tompary, Zhou, & Davachi, 2020). Consistent with this theory, prior work in long-term memory consolidation, which examines gist memory that is newly acquired, has shown that over time, as the strength of gist memory is preserved or improves and/or that of item memory decreases, items that are consistent with the gist are recalled more precisely (Richter, Bays, Jeyarathnarajah, & Simons, 2019; Sweegers & Talamini, 2014; Tompary, Zhou, & Davachi, 2020). However, these results did not demonstrate a gist-based bias, a distortion of item memory from such a newly acquired gist. An increasing bias of item memories towards the remembered gist — in this paradigm, the reported center — would be strong evidence for the increasing strength of gist memory over item memories. Our interest in examining the influence of gist memory on item memory at long delays stems from a desire to bridge the literature reviewed above with a potentially related literature reporting the effects of prior knowledge on memory retrieval (e.g., Huttenlocher, Hedges, & Vevea, 2000; Tompary & Thompson-Schill, 2021).

The current study aimed to understand the relation between item and gist memory over the course of a month. In Experiment 1, we trained three groups of participants on spatial locations of six landmarks, and we measured the change of error in memory of these items as well as the memory for the gist (i.e., the center participants reported) at one of three delay periods: 24 hours, one week, or one month. We predicted that the accuracy of the reported center would persist or improve despite the accuracy of retrieved items decreasing

over a month, as seen in prior work. We extended prior observations by including two new measures—estimated center and gist-based bias—in order to explore how the relation between memories for items and the gist changes over the course of one month. In order to understand how influence of a gist memory on item memory changes over time, we compared observed bias from participants with bias generated under two simulations, one assuming that participants only had item memory, the other assuming that participants had item memory and a separate gist representation.

In Experiment 2, we explored the influence of an "outlier" item in spatial location both on the gist, and on the relation between item and gist memories demonstrated in Experiment 1. Research in long-term memory (Richards et al., 2014) and working memory (Whitney & Leib, 2018) shows that outliers that are inconsistent with the pattern across all items differently influence the memory for the pattern, compared to items that are more consistent with the pattern. Outliers greatly disrupted or shifted the overall pattern (Richards et al., 2014), or were discounted in estimating the pattern (Haberman & Whitney, 2010) compared to other items. In Experiment 2 we examined the extent to which the gist representation was influenced more or less by an outlier item over time, as well as whether bias in item memory was to the center location including or excluding the outlier item with the same simulation approach as in Experiment 1. Taken together, these experiments provide new information about how item and gist memory are consolidated over time.

Results

Experiment 1 Item memories were operationalized as "landmarks" (i.e., dots associated with unique landmark names) in six locations on a laptop screen (Fig. 1). In Session 1, 130

participants learned the locations of the landmarks individually through a training to criterion procedure (see Fig. 1 and Methods for details). After training, participants were tested on item and gist memory: First, they indicated their guess about the center of the landmarks (gist memory test), and then they recalled each landmark location, without feedback and in a random order (item memory test). After 24 hours (n = 44), one week (n = 43), or one month (n = 43), participants returned for Session 2, during which they completed the gist memory test followed by the item memory test again. This testing order was chosen to reduce the influence of item memories on reported gist.

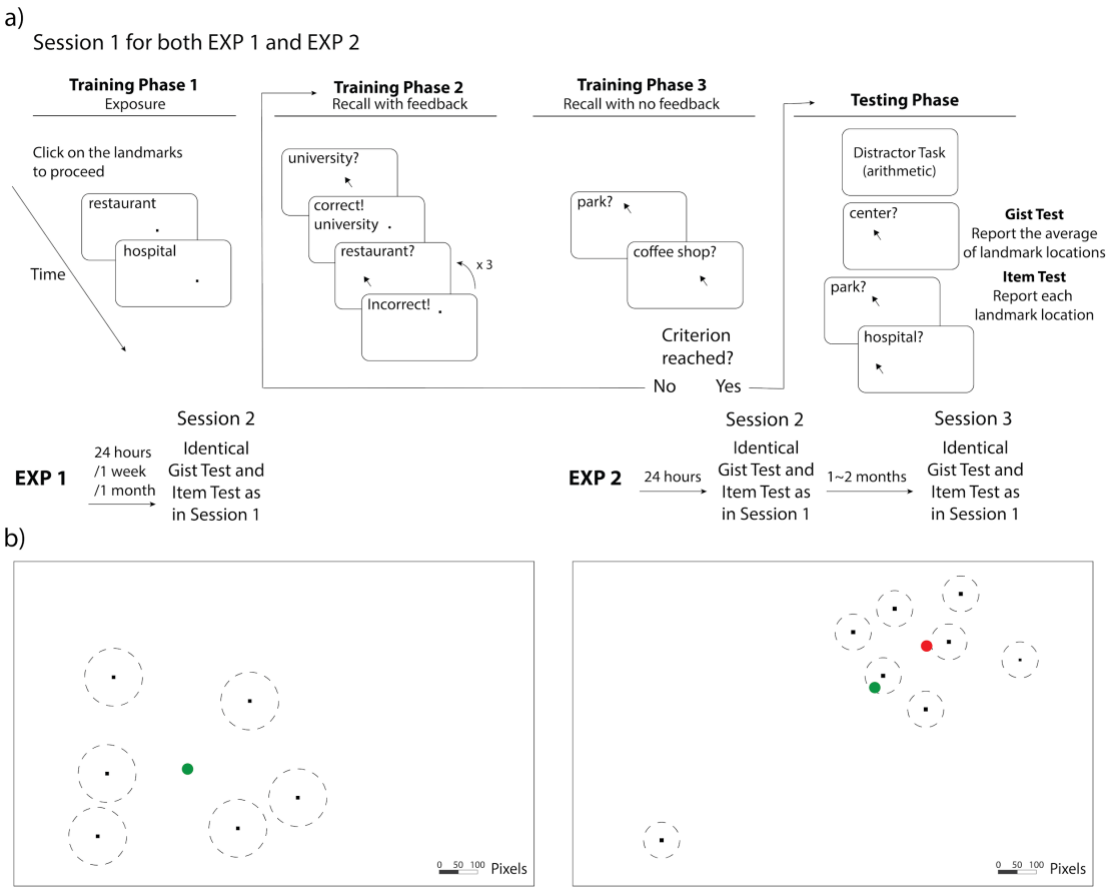


Fig. 2.1. **a)** Schematic illustration of the procedure for Experiment 1 and 2. The procedure of Session 1 is the same for Experiment 1 and 2 (with the exception of the number of trials). Participants completed cycles of encoding (with feedback) and evaluation (without feedback) until they could retrieve each landmark individually within the training criteria. **b)** An illustration of the location of the stimuli (drawn to scale) for Experiment 1 and 2. The locations (black dots) were the same for all participants, but the mapping between the location and landmark name was randomized for each participant. The dash lines around the dots indicate the training criteria (80 pixels for Experiment 1 and 50 pixels for Experiment 2). The green circle indicates the center of these encoded locations and the red circle indicates the 'local' center of the encoded locations (excluding the outlier) in Experiment 2.

Gist memory decreased less than item memory over time

We developed an error measurement for the accuracy of item and gist memory (Fig. 2a; See Methods for details). All three delay groups performed above chance on both the item memory test (compared to chance defined as the average of distance between encoded item locations and center of the screen, and compared to chance defined as the average of distance between each encoded item location and center of all encoded locations) and the gist memory test (chance defined as the distance between the center of encoded locations and center of the screen) at both Session 1 and 2 (all $p < .0001$; Fig 2.8). To examine how item and gist memory changed over time, we conducted a 3 (group: 24-hour, 1-week, and 1-month) X 2 (memory type: item, center) aligned ranks transformation ANOVA of the difference in error between Session 1 and Session 2 (because the data were not normally

distributed). This test revealed a main effect of group, $F(2, 254) = 42.26$, $p < .001$, a main effect of memory type, $F(1, 254) = 99.36$, $p < .001$, and an interaction between group and memory type, $F(2, 254) = 23.76$, $p < .001$. This interaction reveals that the error in retrieved items increased more over time compared to error in the reported center (Fig. 2b). Specifically, whereas each pairwise comparison between groups was significant for item memory (Mann-Whitney tests: all p's $< .01$), the only reliable group difference for gist memory was between the 24-hour and 1-month groups ($U = 685$, $p = .026$). In addition, the change in retrieval error of item locations was significantly higher than that of reported center at a delay of 24 hours (Wilcoxon signed rank tests: $Z = 2.14$, $p = .03$), one week ($Z = 5.79$, $p < .001$), and one month ($Z = 6.53$, $p < .0001$). These results showed that item memories decreased significantly more relative to gist memory over time.

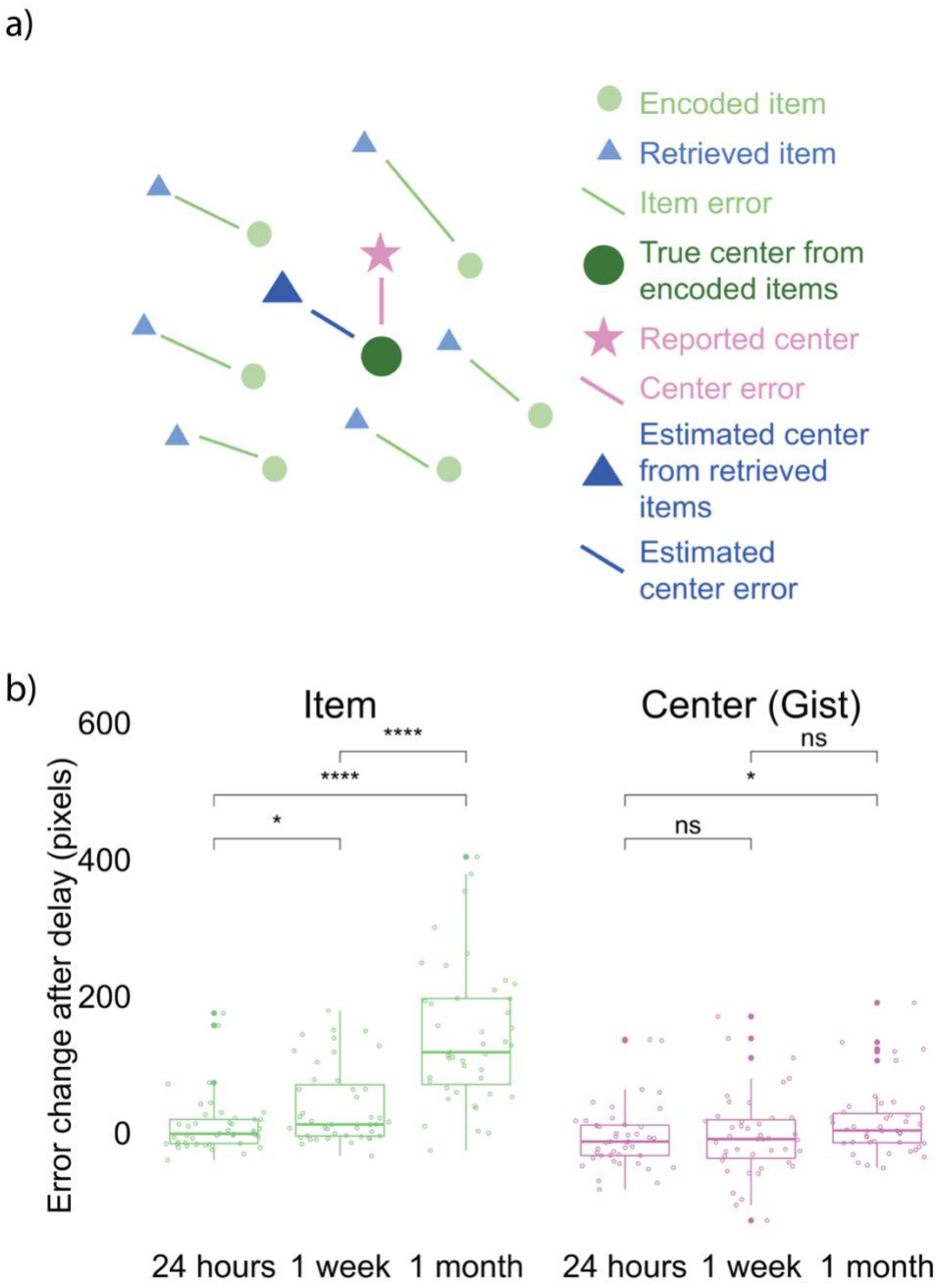


Fig. 2.2. Error measurements and results. **a)** Error measurements. **b)** Change in error by group and memory type (the band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points

indicate outliers). Greater values indicate an increase in error in Session 2 over Session 1. * indicates $p < .05$ and **** $p < .0001$ by Mann-Whitney tests. Fig 2.8 shows the absolute error for both item and gist memory at Session 1 and 2. Figure 2.9 shows the error change over time in reported gist, estimated gist, and simulated gist based on a simple item-only simulation (discussed at the end of Experiment 1 result section).

Positive relationship between item and gist memories across one month

To explore the relation between item and gist memory over time, we used a linear model to evaluate the effects of estimated center error, delay group, and their interaction on reported center error. We found that estimated center error significantly predicted reported center error (SSE = 48138, $F(1, 124) = 18.31$, $p < .001$). The effect of delay (SSE = 4873, $F(2, 124) = 0.93$, $p = .40$) and the interaction between the estimated center error and delay (SSE = 11729, $F(2, 124) = 2.23$, $p = .11$) on reported center error was not significant. These results indicate a stable relation between item and gist memory over time; our subsequent analyses will examine the source of this relation.

Item memory retrieval biased towards gist over time

The positive correlation could indicate that participants' reported gist was still supported by individual item memories, or alternatively that the reported gist was influencing the retrieval of items. To examine the direction of the relation between item and gist memories, we developed a bias measurement (see Methods for details) as an index of how much the retrieval of each item memory is biased towards the gist representation (Fig. 2.3a). We compared the observed bias computed from data collected from participants at each time

point with an *item-only* simulated bias, which assumed learners only had item memory, such that the magnitude of error for each simulated retrieved location would be the same as the corresponding item memory collected from participants, but the direction of the simulated location would be random. We also compared the observed bias to an item-plus-gist simulated bias which assumed both item memory and an additional influence from the gist memory (abbreviated as *gist simulated bias* below for simplicity), such that each simulated retrieved location would be generated from the same error as in the item-only simulation but the probability of a retrieved location being simulated is weighted by its distance towards the reported center. Therefore, a location that is closer to the center will have higher probability of being retrieved in the simulation (see Methods for details). We used the reported center (rather than the true center of encoded items) in this analysis because we found a decrease in accuracy of the reported center after one month compared to 24 hours, as discussed above. In the next section "Follow-up bias analyses", we repeated the bias analysis using the true center of encoded items, for consistency with common practices in ensemble perception research (Brady & Alvarez, 2011; Lew & Vul, 2015).

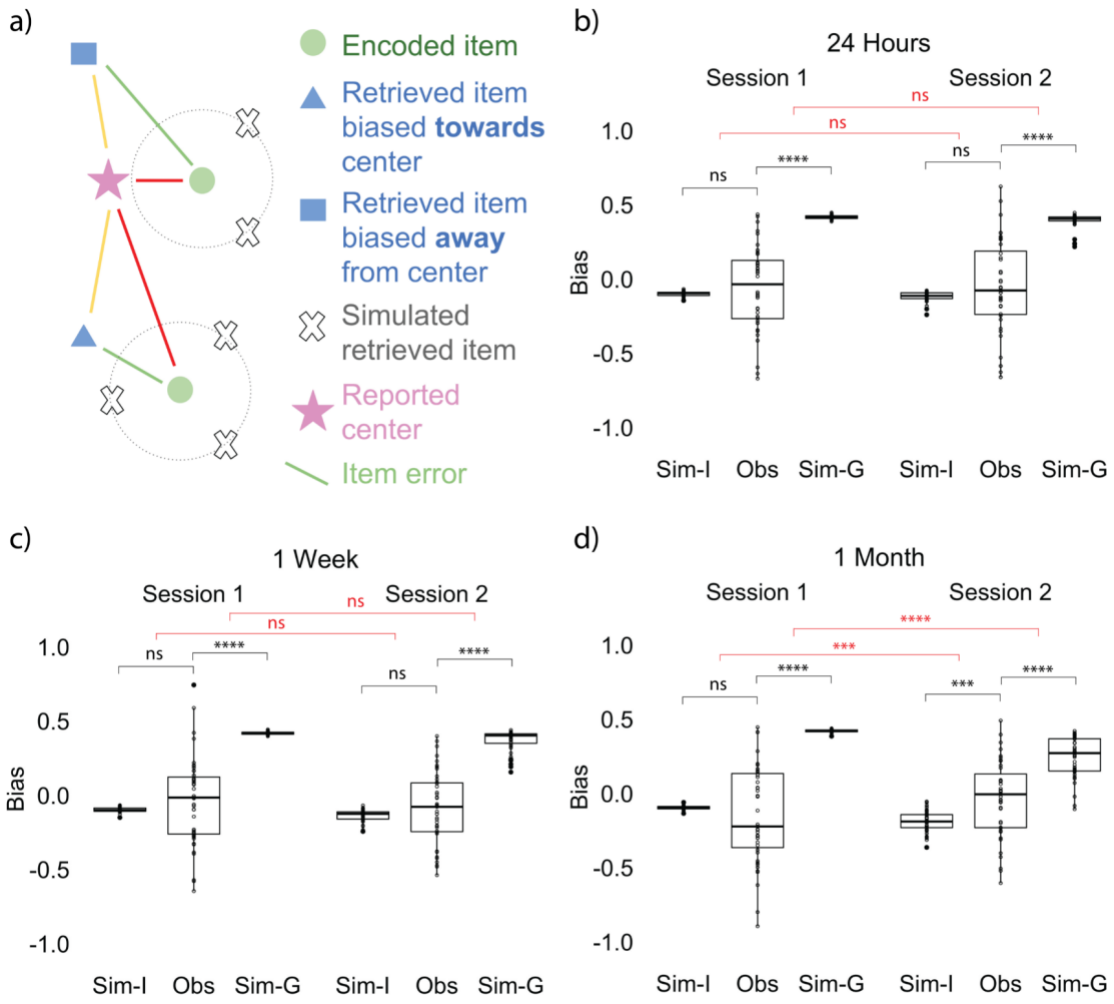


Fig. 2.3. a) Bias measurement. The bias for each recalled location is (red - yellow) / green. The blue square is an example of a recalled item that is biased away from the reported center and the blue triangle is an example of a recalled item that is biased towards the reported center. Bias for each participant is an average of bias for all the locations. **b**, c, d) Item-only simulated bias (Sim-I), observed bias (Obs), and gist simulated bias (Sim-G) at each session for delay groups of 24 hours, 1 week, and 1 month (the band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers). * indicates *p*

< .05, ** p < .01, *** p < .001, and **** p < .0001 by t-tests between observed bias and simulated biases (black) and t-tests comparing the difference in observed bias and simulated biases between sessions (red).

In order to examine whether participants' item memory became more dissimilar over time to what would be expected from having item representations and no separate gist representation, we conducted a 2 (session: Session 1 vs. Session 2) x 3 (delay groups: 24 hours, 1 week, and 1 month) ANOVA for the difference between the observed bias and item-only simulated bias. This test revealed a significant interaction between delay group and session, $F(2, 254) = 3.53$, $p = 0.03$, indicating that the difference between item-only simulated bias and the observed bias significantly increased over time (Fig 2.3). Follow-up t-tests for all delay groups revealed that for the one month group only, the difference between the observed bias and the item-only simulated data was significantly higher for Session 2 compared to Session 1, $t(80.76) = 3.18$, $p < .01$. No across-session comparisons for any other delay groups were significant (ps > .65). Furthermore, only the observed bias at 1 month was significantly greater than the item-only simulated bias, $t(42) = 3.73$, $p < .001$ (Fig 2.3b).

In order to examine whether participants' item memory became more similar over time to what would be expected from having both item and gist representations, we conducted an analogous 2 x 3 ANOVA for the difference between gist simulated bias and the observed bias. This test revealed a significant interaction between delay group and session, $F(2, 254) = 6.33$, $p < 0.01$, indicating that the difference between gist simulated bias and the observed bias significantly decreased over time. Follow-up t-tests for all delay groups revealed that

for the one month group only, the difference between the observed bias and the gist simulated bias was significantly lower for Session 2 compared to Session 1, $t(82.50) = 4.39$, $p < .001$. No across-session comparisons for any other delay groups were significant (ps > .65). (Note. Because the level of gist influence added to the gist simulated bias was arbitrarily selected, we did not expect the endpoint value for the observed bias to match the gist simulated bias; we will return to this point in the Discussion.). These results indicate that participants' biases were increasingly consistent with the assumption of a separate gist representation and increasingly inconsistent with reliance only on item memories. By one month (but not after one day or one week), item memory retrieval was biased towards the reported gist[2].

Taken together, the results of Experiment 1 reveal that although there is a persistent relation between item and gist memory during memory consolidation, the nature of this relation changes over time. We suggest that early in memory consolidation, retrieval of gist depends on the successful retrieval of individual items, but then, as item memory weakens over time, the relatively stronger gist memory begins to guide retrieval of item memory.

[2] We used this bias analysis to measure the influence of the gist—as opposed to comparing the distance between the reported center and estimated center or their error difference over time—for two reasons. First, our measure of bias shows the magnitude of the particular direction the items were attracted to. Second, the estimated center was calculated from an aggregation of item memories, which may already have been influenced by the center of these items at a delay (as shown by the bias analysis). Therefore, comparing the estimated center and the reported center over time would not allow us to isolate the influence of the center. To demonstrate this point, we conducted a simulation for the estimated center error, assuming the magnitude of error for each item memory would remain the same but the direction of error would not be systematically influenced by the gist. Such simulated estimated center error is significantly different from participants' reported center error and estimated center error, which suggests that the estimated center computed from the item memories was influenced by the center (Fig 2.9).

As a consequence, this new gist representation can exert influence over memories in ways described by reconstructive memory theories.

Follow-up bias analyses

In order to test factors that may influence the gist-based bias and its generalizability, we conducted the bias analyses as discussed above with the three following modifications (see Methods for additional details): First, to be consistent with common practices in ensemble perception research (Brady & Alvarez, 2011; Lew & Vul, 2015), we repeated the bias analysis using the center of encoded items (big green circle in Fig 2.2a) instead of the reported center. Second, the assumption that each item is weighted the same may be overly simplistic and the weight of the items may influence the representation of the gist and the bias results. Therefore, we computed a center that was a weighted average based on the accuracy of each item, such that items with higher recalled accuracy would be weighted more in the computed center compared to items with lower accuracy. We repeated the bias analysis using this weighted center. Third, the mental representation of the locations that participants encoded may not be a linear transformation of the actual item locations on the computer screen, and this nonlinearity may account for the observed biases in location memory. To capture the potential non-linear warping of the stimulus space, we generalized the Euclidean error measure to a Minkowski's measure, where error $d(a,b) = \sqrt[1.5]{(a_1 - b_1)^{1.5} + (a_2 - b_2)^{1.5}}$, and conducted the same bias analysis.

Across the three analyses, we found the same pattern: The observed bias became more and more dissimilar from the item-only simulated bias, indicated by a significant interaction between delay group and session in the three ANOVAs (all $Fs > 6.42$, $ps < .01$), and became more and more similar to the gist simulated bias over time, indicated by a significant interaction between delay group and session in ANOVAs (all $Fs > 6.81$, $ps < .01$). Furthermore, the observed bias at one month differed from the item-only simulated bias (all $ts > 3.85$, $ps < 001$), but not at other delays. In addition, the bias computed under these three approaches are highly correlated with the bias with reported center in the prior section (all $rs > 0.9$, $ps < .001$), suggesting the varied approaches generated similar bias to the bias in the prior section. In summary, the result of increasing gist-based bias over time replicates in analyses using the center of encoded locations, weighted center based on item accuracy, and with a Minkowski's measure in non-Euclidean space.

Experiment 2 The stimuli and procedure of Experiment 2 (Fig. 2.1) were similar to those of Experiment 1 but differed in two major ways (see Methods for more details). Firstly, we used a repeated measures design in Experiment 2 so that we could observe changes in memory at short (one day) and long (1-2 month) retention intervals within each subject (N = 43). Secondly, one of the landmarks was an "outlier", meaning that its location fell far out of the range of the cluster where the majority of the landmarks were (see Experiment 2 item locations in Fig. 2.1). The inclusion of an outlier location enabled us to examine the influence that a single "atypical" item has not only on the initial estimation of the center (as in Haberman & Whitney, 2010; Richards et al., 2014; Whitney & Leib, 2018) but also on the source of bias in item memory at long delays. In addition, in Session 1, item memory

was derived from the item memory test in the last round of evaluation during training (see the procedure for Experiment 2 in Fig. 2.1) to streamline the session.

Gist memory decreased less than item memory over time

In Experiment 2, we used the same error measurement for the accuracy of item and gist memory as in Experiment 1 (Fig. 2.2a; see Methods for details). Participants performed above chance in both item (i.e., the average of distance between encoded item locations and center of the screen; the average of distance between encoded item locations and center of encoded item locations) and gist memory (i.e., the distance between the center of encoded locations and center of the screen) tests at all sessions (all $p < .0001$; Fig 2.10). To examine how item and gist memory changed over time, we conducted a 2 (delay: short (24 hours) or long (1-2 months) x 2 (memory type: item, center) aligned ranks transformation ANOVA with repeated measures of error change. This test revealed a main effect of delay, $F(1, 126) = 16.20$, $p < .001$, memory type, $F(1, 126) = 68.96$, $p < .001$, and an interaction between delay and memory type, $F(1, 126) = 27.40$, $p < .001$. This interaction indicates that the error of individual item retrieval increased more over time compared to the reported center (Fig. 2.4a). Specifically, whereas item memory error change was higher after 24-hour compared to one to two months by Wilcoxon signed rank test ($Z = 5.28$, $p < .0001$), no such significant difference was detected for gist memory error change ($Z = 0.72$, $p = .47$). Experiment 2 thus replicated the finding observed in Experiment 1 that item memories decreased significantly more relative to gist memory over time.

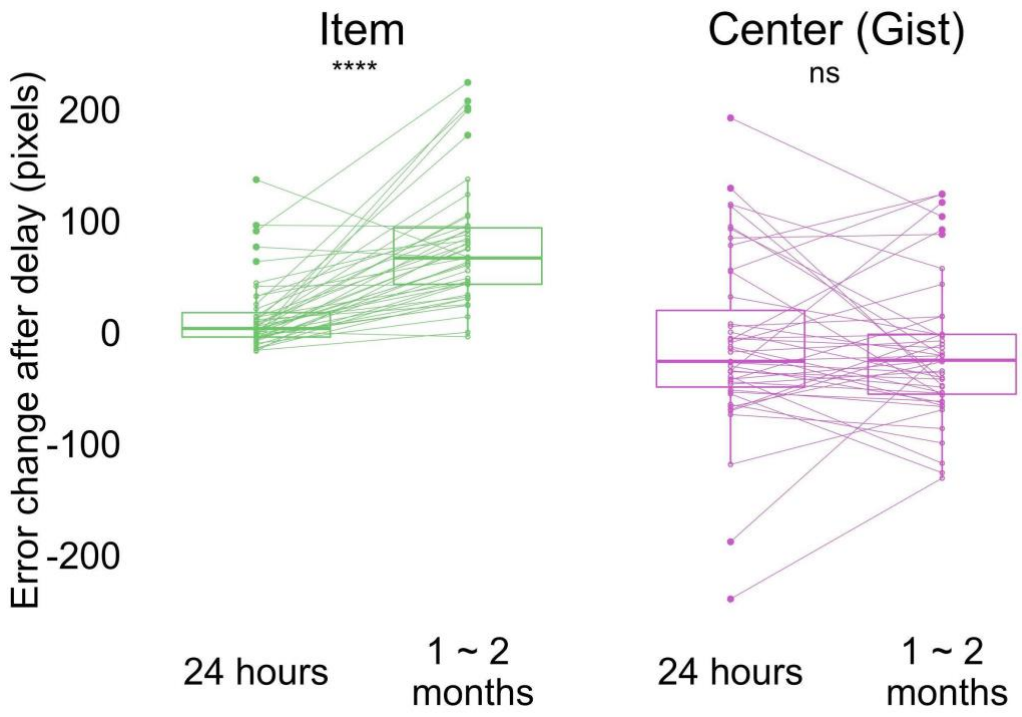


Fig. 2.4. Change in error by delay and memory type (the band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers). Greater values indicate an increase in error from Session 1 after delay. **** $p < .0001$ by Wilcoxon signed rank tests. Dots and lines indicate participants. Fig 2.10 shows the absolute error for both item and memory at all sessions.

Positive relationship between item and gist memories across time

To explore the relation between item and gist memory, we fit a linear mixed effects model on reported center error with fixed effects of delay (24 hours and 1 to 2 months), estimated center error, and their interaction with a random effect of participant to account for repeated measures within participants. We found a significant effect of estimated center

error (*SSE* = 29114.3, $F(1, 69.92) = 12.08$, $p < .001$), but not a main effect of delay (*SSE* = 1420.9, $F(1, 59.59) = 0.59$, $p = .45$) or an interaction between the estimated center error and delay (*SSE* = 1102.9, $F(1, 68.80) = 0.46$, $p = .50$). The result of a persistent relationship between estimated center error and reported center error at short and long retention intervals was replicated in a within-participants design.

Item memory retrieval was biased towards the local gist over time

To examine whether the influence of gist on item memories changes over time, we applied the same bias analysis as in Experiment 1, using participants' reported center of all the retrieved items as bias center (Fig. 2.1b). We conducted a one-way repeated measures ANOVA on the difference between the observed bias and item-only simulated bias across sessions (after training, 24 hours, and 1 month) and also the same ANOVA on the difference between the observed bias and gist simulated bias. In contrast to Experiment 1, the difference between item-only simulated bias and the observed bias did not significantly change over time, $F(2, 84) = 1.34$, $p = 0.28$ (Fig. 2.5a). Furthermore, unlike in Experiment 1, the observed bias was not significantly higher compared to the item-only simulated bias at long retention, $t(42) = -0.47$, $p = .64$ (Fig. 2.5a). At the same time, the difference between gist simulated bias and the observed bias only marginally decreased over time, $F(2, 84) = 3.04$, $p = 0.05$ (Fig. 2.5a). This may be driven by an unpredicted negative bias (i.e., bias away from the reported gist) immediately after learning, $t(42) = -2.09$, $p = .04$, and after a short retention interval, $t(42) = -2.27$, $p = .03$, revealed by t-tests against the item-only simulation.

What might explain the different bias results between Experiments 1 and 2? We suspect this is the result of the outlier item. Prior work in visual working memory research showed that outliers were discounted in estimating the gist (Haberman & Whitney, 2010). In our Experiment 1, where there was not an outlier, we saw that the item retrieval was biased towards the center of all of the items; however, in Experiment 2, the center of *most* of the items would be the local center excluding the outlier (Fig 2.1b). It is possible that for participants in Experiment 2, the items were biased towards the local clustering center excluding the outlier.

In order to test this possibility, we conducted an analysis that computed the gist-based bias of the items using the local center (the true center from the encoded items disregarding the outlier). As in Experiment 1, the difference between item-only simulated bias and observed bias significantly increased over time, $F(2, 84) = 8.51$, $p < 0.001$ (Fig. 2.5b). Follow-up paired t-tests showed that the difference between item-only simulated bias and observed bias at long retention was significantly higher compared to the difference after training, $t(42) = -3.67$, $p < .001$, and compared to the difference at short retention, $t(42) = -2.87$, $p < .01$). The same comparison between after training and 24 hours was not significant, $t(42) = -1.49$, $p = .14$).

On the other hand, the difference between gist simulated bias and observed bias decreased over time, $F(2, 84) = 13.80$, $p < 0.001$ (Fig. 2.5b). Follow-up paired t-tests showed that the difference between gist simulated bias and observed bias at long retention was significantly smaller than the difference after training, $t(42) = -4.59$, $p < .001$, and compared to the difference at short retention, $t(42) = -3.81$, $p < .001$. The same comparison between after training and 24 hours was not significant, $t(42) = -1.62$, $p = .11$. Furthermore,

only the observed bias at the long retention interval was significantly greater than the item-only simulated bias ($t(42) = 2.28$, $p = .03$; Fig. 2.5b). This increased bias was observed even for the outlier item: Retrieval of the location of the outlier item was significantly more biased towards the local center after a long retention interval compared to a short retention interval, revealed by a comparison to the item-only simulated bias ($Z = 2.46$, $p = .01$). These results indicate that item memories in Experiment 2 were biased, at long retention intervals, towards the center as in Experiment 1, but that the "center" in Experiment 2 was not the global center but instead the local center excluding the outlier item.

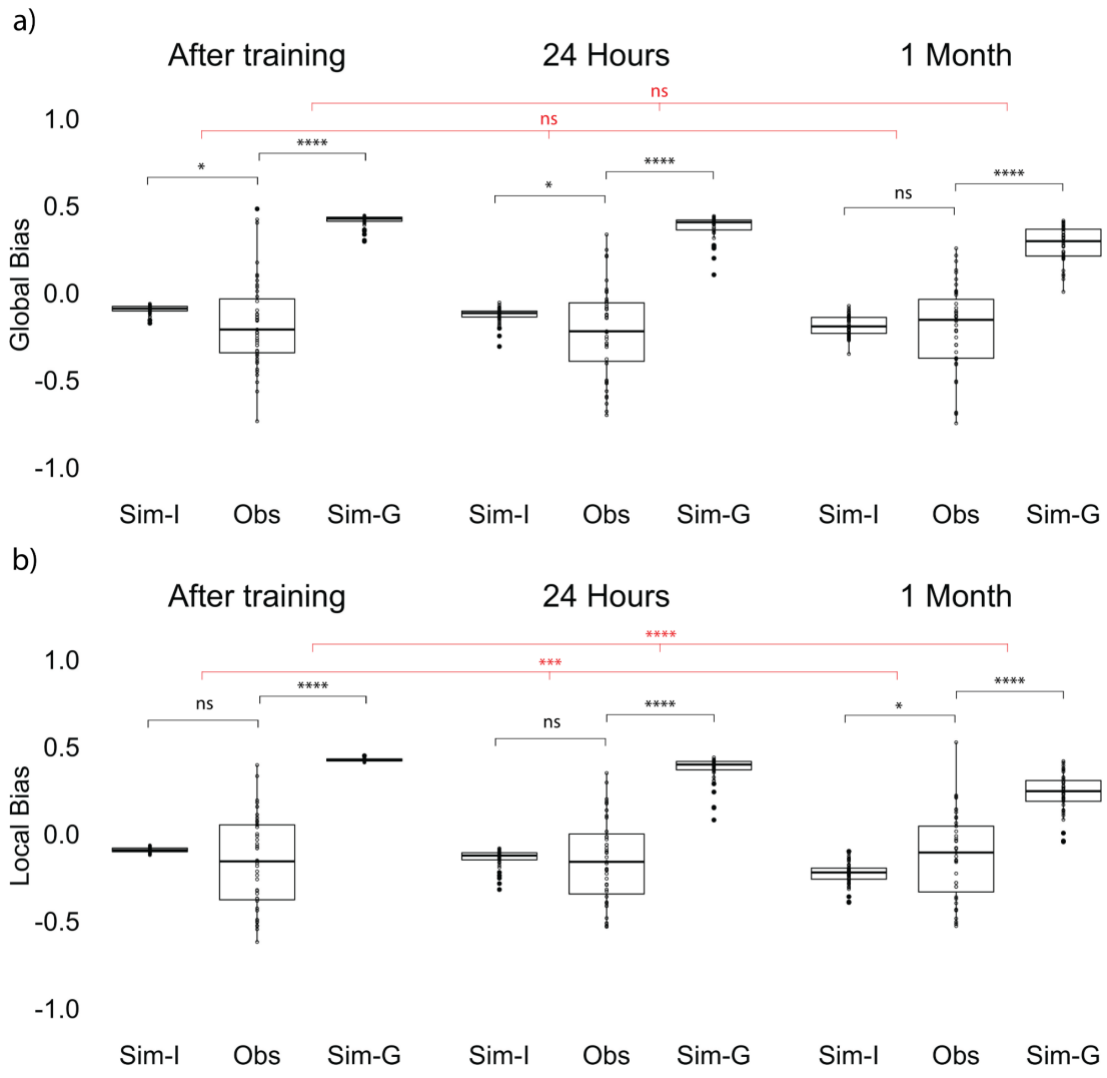


Fig. 2.5. a) Global observed bias (Obs), item-only simulated bias (Sim-I), and gist simulated bias (Sim-G) at each session. Global bias uses the report center. b) Local observed bias (Obs), item-only simulated bias (Sim-I), and gist simulated bias (Sim-G) at each session. Local bias excludes the outlier item when estimating the center. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers. * indicates $p < .05$, *** $p < .001$, and **** $p < .0001$ by paired t-tests between observed data and

simulation (black) and repeated measures ANOVA comparing the difference in data and simulations across sessions (red).

Over-weighting of the outlier in gist memory

Our analysis of item bias suggests that the outlier is "discarded" as a member of the cluster of locations, which is consistent with some prior studies (e.g., Haberman & Whitney, 2010); however, other work has shown that outliers can greatly disrupt or shift the representation of a set of events (Richards et al., 2014). Could both be happening in this paradigm? In order to explore the influence of the outlier on the representation of gist, we applied a weighted model adapted from working memory literature and computed an estimation of the weight of the outlier in the reported gist (Haberman & Whitney, 2010; see Methods for details).

The estimated weight of the outlier was significantly higher than 0.125 (i.e. the level assuming equal weights across all items) immediately after learning, $t(42) = 2.14$, $p = .04$, after a short retention interval, $t(42) = 2.89$, $p < .01$, and after a long retention interval, $t(42) = 3.83$, $p < .001$, (Fig. 2.6). The change in outlier weight after short compared to long retention intervals did not significantly differ ($t(42) = 0.92$, $p = .36$). In other words, the outlier has not been discarded from the set, but quite to the contrary, the outlier has a disproportionate influence on the explicit retrieval of gist after a delay. In contrast, the implicit effect of the center on bias in item retrieval seems to emerge from a center that is uninfluenced by the outlier (Fig. 2.5b). These results revealed that the outlier consistently influenced participants' reported center more than other items at all tested time points.

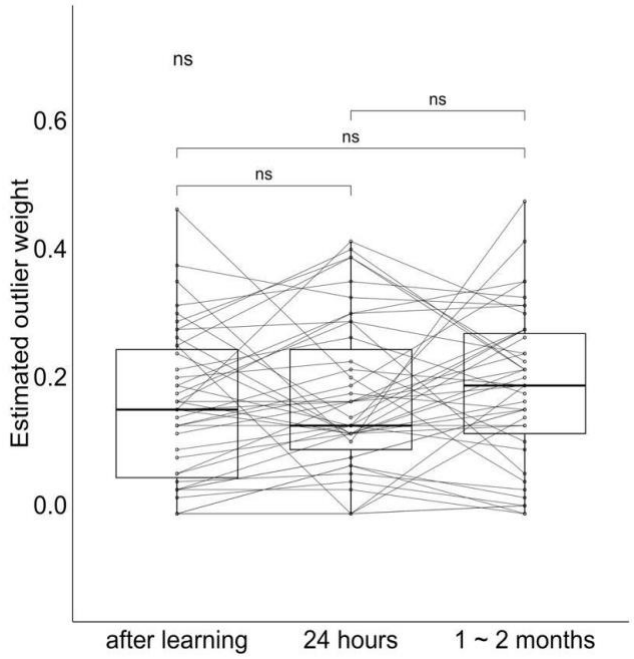


Fig. 2.6. Outlier weight values at each session. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range. Dots and lines indicate participants.

Taken together, Experiment 2 replicated the main findings from Experiment 1 that gist memory decreased less compared to item memories over time (Fig. 2.4) and a positive relationship between item and gist memories in a within-subject design. The "outlier" item changed the relationship between items and reported global center after a long retention interval. By one to two months, items were no longer biased towards the global reported center which overweighted the outlier. Instead, they were biased towards the local center excluding the outlier over time.

Discussion

We examined how human learners extract the "gist" (generalities, common properties, or summary statistics) across individual instances, and how memory for these instances and for the gist evolve and influence each other over time through two behavioral experiments spanning one to two months. We demonstrated that the accuracy of item memory (memory for spatial locations on a screen) decreased more compared to the accuracy of gist memory (center of the locations) over time, though there was a persistent positive correlation between them. Critically, item memories were increasingly biased towards the gist over time. Participants' biases grew less similar over time to a simulation relying only on memory for individual items and more similar to a simulation assuming a separate gist representation. In the presence of an outlier item, the local gist, excluding the outlier, became the source of bias, instead of the gist participants directly reported, which consistently overweighted the outlier across time. We think that gist memories, initially built from item memory, gradually developed to guide item memory as their relative strength changed over time.

Consistent with prior research (Antony et al., 2020; Berens et al., 2020; Lutz, Diekelmann, Hinse-Stern, Born, & Rauss, 2017; Posner & Keele, 1970), item memory became less accurate over time while gist memory remained relatively intact over time. Our findings converge with this prior research even when using an explicit instruction to retrieve gist memory, rather than inferring gist from another measure as in prior research. A shortcoming of prior research is that the relation between item and gist memory over time is rarely assessed. We showed that the relationship between item memory and gist memory persisted across delay periods despite decreased accuracy in item memory. This

relationship could have resulted from the influence of gist memory on the retrieval of item memory, from the influence of item memory on gist memory, or both. Our gist-based bias results shed light on the direction of this relationship: Item memory retrieval was biased towards gist only after one month, which suggested that the correlation at one month was likely to be due to the influence of gist memory on the retrieval of item memory.

Our findings that items are increasingly biased towards the gist as the accuracy of item memories decreases over time provide new evidence for the memory reconstruction framework, which proposes that memory retrieval is a combination of different sources with varying strength (Brady et al., 2015; Hemmer & Steyvers, 2009; Huttenlocher et al., 2000; Tompary et al., 2020). Our work extends prior evidence of increased schematization in memory consolidation (Richards et al., 2014; Richter et al., 2019; Tompary et al., 2020) by demonstrating a new form of influence from the gist on item memory: gist-based bias. In contrast to prior memory consolidation research that showed increased schematization earlier than one month (Graves et al., Richter et al., 2019; Tompary et al., 2020), the gist-based bias in our current work did not increase by 24 hours or one week. This discrepancy could be because the intensive training participants experienced in our paradigm increased the strength of item memories relative to gist memory during learning, and only after a long retention interval did the strength of item memory decrease to an extent that allowed bias to manifest.

The increased bias may reflect a slow systems consolidation process that results in a qualitatively different memory representation after longer retention intervals (Richards et al., 2014). An increased reliance on neocortical areas over time would be expected to strengthen gist memory, to the extent that neocortex tends to represent information in a

'semanticized' form (Sekeres, Moscovitch, & Winocur, 2017). The results are also consistent, however, with a change in reliance on different forms of memory within the same memory systems. The current results are not diagnostic on this point — they are consistent with a range of theories on the interplay between episodic and semantic memories over time (Renoult, Irish, Moscovitch & Rugg, 2019; Richards et al., 2014; Robin & Moscovitch, 2017; Winocur & Moscovitch, 2011; Sekeres, Winocur, & Moscovitch, 2018).

Our results of increasing gist-based bias over time parallel visual working memory work, which shows evidence of a hierarchical organization of memory: items are more biased towards their center as uncertainty increases in order to increase the overall precision of retrieval (Brady & Alvarez, 2011; Lew & Vul, 2015; Orhan & Jacobs, 2013). Our results detected a similar gist-based bias in long-term memory consolidation. Moreover, in Experiment 2, after a long retention interval, the reported gist overweighted the outlier, whereas the item memories were biased towards the local gist which discounted the outlier. This finding also mirrors prior ensemble perception results that outliers are discounted or excluded in estimating summary statistics (De Gardelle & Summerfield, 2011) and suggests that the gist influencing item retrieval is not a simple average of the items. The results might reveal two different sampling strategies for gist extraction. Because participants had explicit knowledge about the outlier, they might have given more weight to the outlier in explicitly recalling and reporting the gist, similar to the change in the pattern by inconsistent items observed in long-term memory work (Richards et al., 2014). In contrast, the local center that influenced the items might reflect an implicit representation with a sampling strategy discounting the outlier, consistent with findings in

perception work (De Gardelle & Summerfield, 2011; Haberman & Whitney, 2010). Our results suggest that visual working memory and long-term memory might be underpinned by a similar reconstructive mechanism and open up new directions to bridge the two fields.

One limitation of the current experiments is that the testing order (i.e., gist memory before item memory) might have encouraged the retrieval of the items to be consistent with the gist (Tversky & Kahneman, 1974; Mutluturk & Boduroglu, 2014). We initially chose this order because we were most interested in the change in gist representation and wanted to minimize the influence of item memories on gist estimation in later recall. We also were concerned that the extent that these two tests influenced the other was not symmetric; in other words, the influence of item memory on gist might be more pronounced than the influence of gist on item memory. Because the testing order is the same for the three different delay intervals, we reasoned that the changes in item memory and bias across delay groups could not simply be a result of the order. In addition, in Experiment 2, the items were not biased towards the center participants reported, suggesting that even if the gist test occurring before item test influences the recall for the items, the influence may be minor. However, multiplicative effects, such as floor or ceiling effects present only at one time point, could still influence the results and the influence from the testing order may still exist. Although the testing order is likely not to influence the change over time, the bias for all delay groups may be lower overall under the reverse testing order. More studies with counterbalanced testing order will be helpful to evaluate this possibility.

Future research can be done to test the generality of our findings to other domains of human cognition. It would be interesting to explore whether our findings, which considered

gist memory as a spatial average, would generalize to a broader definition of gist, such as gist-like memory for events (Moscovitch, Cabeza, Winocur, Nadel, 2016). For example, when first learning what a birthday party is from attending a few, the "gist" representation of a birthday party may be dominated by memory for a few parties, but over time the gist becomes a more stable representation that can influence retrieval of those specific birthday party events. In addition, the dissociation of gist-based bias in Experiment 2 also mirrors the dissociable implicit and explicit attitude in social categories (Gawronski & Bodenhausen, 2006). More work could further disentangle these processes in long-term memory consolidation, which could enlighten our understanding of the cognitive mechanism underlying the formation of gist in social categories.

We began by posing the question of how one extracts a summary statistic from individual instances, but without a doubt, the summary statistic we have used here to answer this question—the arithmetic average of x,y coordinates—is overly simplistic. Firstly, the item-only simulations in our work are oversimplified compared to the item-only models in the categorization literature (Nosofsky, 1988). There surely could be other more sophisticated item-only models that can fit our data. However, our results put a new constraint (i.e., a gist-based bias) for item-only models in long-term memory. Secondly, our implementation of a gist representation in our bias simulations was very simplistic. For example, we assumed an arbitrary amount of gist influence, implementing a more qualitative than quantitative assessment of the presence of a gist representation. Additional experiments with more within-subject statistical power could be used to constrain models that quantify the precise amount of gist influence (as a parameter in individual model fits). Thirdly, this gist influence may be influenced by many other factors, such as the variability

in item locations, the accuracy and the confidence of item memories, the distance from the items to the center or to the screen boundary (Intraub & Richardson, 1989), individual differences in cognitive functions (e.g., executive control), and the demand characteristics of explicitly recalling the center. The current design did not allow for enough variability to tease apart these possibilities, but future research systematically manipulating these factors will be helpful in addressing these issues. Finally, and perhaps most importantly, the principles that govern aggregation of individual spatial locations do not in any obvious way translate to the nature of summary statistics for other episodic memories (like that average bear!). Although there is much work to be done to understand the ways in which we aggregate information across multiple experiences, the current experiments should provide a useful launching off pad for future explorations of this question.

In summary, we have shown that memory for individual items and memory for the gist of a set of items changed over the course of long-term memory consolidation. We propose that the gist that was initially extracted from item memories gradually started to guide item memory retrieval over longer durations as their relative memory strength changed. These findings bridge research in areas of cognitive science ranging from perception and working memory to episodic and semantic memory, providing important new insights into our ability both to learn about distinct events and to generalize across similar experiences.

Methods

Experiment 1.

Participants. In Experiment 1, we recruited 147 members of the University of Pennsylvania community (18-30 years old; normal or corrected to normal vision) to

participate in the experiment for monetary compensation. Participants selected to sign up for a second session that followed their first session by either 24 hours, one week, or one month[3]. Sample size was based on Experiment 2 which was conducted first[4]. We excluded 10 participants because of low performance on Session 1 (i.e., reported gist was out of the scope of the learned landmarks) and then 7 participants because of individual and gist performance of any sessions lower than 3 *SD* below average. Our reported results thus include 130 participants, with 44 participants in the 24-hours group (age: M = 21.3, SD = 2.9, gender: 61% females), 43 participants in the one-week group (age: M = 21.9, SD = 2.6, gender: 67% females), and 43 participants in the one-month group (age: M = 21.4, SD = 2.0, gender: 74% females). All procedures were approved by University of Pennsylvania IRB (IRB #705915, Linguistic and Nonlinguistic Functions of Frontal Cortex).

Procedure. The experimental procedure is displayed in Fig. 2.1. All participants completed Sessions 1 and 2; the only difference between groups was the time delay between sessions. Session 1 included training and testing. During training, participants were trained to retrieve six landmark locations consecutively on a laptop until their retrieval error for each landmark was fewer than 80 pixels in any direction. 80 pixels was chosen to be the criterion because it was less than 1/2 of the shortest distance between any pairs of the encoded locations, and thus would ensure that participants could differentiate the locations in recall.

[3] Because participants were not randomly assigned into three different delay conditions, a difference in expectation may influence their learning and consolidation. We did not find evidence, however, for any differences in behavior between groups at initial learning (Fig 2.8). Also, the results from Experiment 1 replicated in Experiment 2 with a within-subject design.

[4] Namely, we set the number of subjects after exclusion in Experiment 2 as a minimum sample size. After we reached the sample size, we continued to recruit participants until the end of the academic term.

The training included three phases. In Phase 1, the landmarks appeared on the screen one at a time, and participants were required to click on each landmark to proceed. Fig 2.1 illustrates the landmark locations; note, on each trial, only one location was presented (never the full map), and the center of the encoded locations was never presented to participants. In Phase 2, we asked participants to recall the location for each landmark by clicking on the screen when given its name as a cue, and we gave them feedback about their guesses: participants had 3 attempts to recall each landmark location. For each attempt, if the distance between the recalled location and retrieved location satisfied the training criterion (i.e., 80 pixels), the correct location would be shown on the screen; otherwise, a message would be prompted that their attempt was incorrect. The correct location would be shown on the screen after three incorrect attempts. In Phase 3, participants recalled each landmark consecutively without feedback, one at a time. If each of the retrieved landmarks fell in the range of 80 pixels, the participant could proceed to testing; if not, the participant was redirected back to Phase 2 to receive more training. After participants reached the training criterion, they completed ten unrelated arithmetic problems, in order to minimize potential influences from working memory. Finally, participants were tested on their memory of the locations: They indicated their guess about the center of the landmarks (gist memory test), with an instruction "Indicate the center (average location) of the landmarks you have seen". Then, they separately recalled each landmark location (item memory test, which was identical to the recall procedure in Phase 3). The order of items was randomized in all phases in training and testing. In both tests, participants were incentivized to be accurate through bonus payments. They would receive a bonus of 1 dollar for the gist memory test if their error was within 100 pixels and a bonus of 1 dollar for the item memory

test if their average error across all items was within 80 pixels. All trials were self-paced. The total time for Session 1 was approximately 12 minutes.

After 24 hours, one week, or one month, participants returned for Session 2. Session 2 was identical to the gist test and item test in Session 1, in which participants first reported the center and then the location of each landmark. Trials were again self-paced. The total time for Session 2 was approximately 5 minutes. Participants could choose to quit the experiment after Session 1 and receive 10 dollars for their time, otherwise they would be paid after Session 2. The payment ranged from 16 to 20 dollars, depending on participants' performance in their gist memory and item memory test.

Error Measurement. In order to measure the accuracy for item memory (memory for each landmark), gist memory (reported memory for the center of the landmarks), and the estimated gist (center of all the retrieved items), we developed three error measurements as follows.

Item Memory Error (green line in Fig. 2.2a): The error for each item was defined as the Euclidean distance between the retrieved location for each landmark and its encoded location, where

d (a, b) = $\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$. Each participant's item memory error was computed as the average error for the six landmarks. Chance performance based on the center of the screen is 348 pixels, which is determined by the average Euclidean distance between the center of the laptop screen and each encoded item location. This distance corresponds to what participants' performance would be if they only remembered the center of the screen and just clicked the center of screen when asked to recall an item.

Mathematically, this distance corresponded to the average error a participant would have if they guessed anywhere on screen. Chance performance based on the center of the screen is 267 pixels, which is determined by the average Euclidean distance between the center of encoded item locations and each encoded item location. This distance corresponds to what participants' performance would be if they only remembered the center of item locations and just clicked that center when asked to recall an item.

Gist Memory Error (purple line in Fig. 2.2a): The error for gist memory was defined as the reported center error, which was the Euclidean distance between the participant's reported center and the true center of all the encoded items. Chance performance is 270 pixels, which is the Euclidean distance between the center of the laptop screen and the true center of all encoded locations. This distance corresponds to what participants' performance would be if they just clicked the center of screen when asked to report the center.

Estimated Gist Memory Error (blue line in Fig. 2.2a): The error for the estimated gist based on items was defined as the estimated center error, which was the Euclidean distance between the center of each participant's retrieved item locations and the true center of all encoded locations. In other words, the estimated center can be thought of as what the participant's gist estimate would be if it were directly computed by averaging across all retrieved item locations.

Bias Measurement. In order to measure the influence of gist on item memory, we developed a bias measurement as follows. Since the error analysis revealed a decrease in gist memory (i.e., reported center) after a month, there could be a difference in using

reported center and using the true center of encoded items as bias center. We initially used the reported center as the center for the bias analysis. However, we also computed a bias using the center of encoded items to be consistent with common practices in ensemble perception research (Brady & Alvarez, 2011; Lew & Vul, 2015).

Observed Bias: The bias towards the center for each retrieved item was defined as the relative difference in distance between a participant's reported center and each landmark's encoded location versus each landmark's retrieved location. This relative difference was then divided by the error for that landmark:

$observed\ bias = \frac{(d(Encoded\ Item, Reported\ Gist) - d(Retrieved\ Item, Reported\ Gist))}{d(Encoded\ Item, Retrieved\ Item)}$ where

d (a, b) = $\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$ (Fig. 2.3b). Bias thus can range between -1 and 1 and bias > 0 indicates that item memory is biased towards the center while bias < 0 indicates that the item is biased away from the center. Each participant's bias not controlling for error was computed as the average across the biases of the 6 landmarks.

Item-only simulation: The item-only simulation assumed that the magnitude of error for each item memory would be the same as the corresponding item memory collected from participants, but the direction of simulated recalled location would not be systematically influenced by the gist. Following this assumption, we generated 1000 simulations for each participant. Each simulation consisted of six simulated retrieved items, corresponding to all the six landmark locations. For each location, we randomly generated a retrieved location based on the participant's true error for this specific location, allowing its angle relative to the encoded location to vary randomly across the simulations (Fig. 2.7b; gray cross). If a simulated location fell outside the boundaries of the screen, the algorithm

generated a new location. The bias value for each of the 1000 simulations was the average across each simulation's six retrieved locations. The item-only simulated bias for each participant was the average across the 1000 simulations.

Item-plus-gist simulation (abbreviated as gist simulation for simplicity): The gist simulation assumed that the magnitude of error for each item memory would be the same as the corresponding item memory collected from participants, but the reported center systematically influences the probability of recalled locations (instead of the uniform probability distribution around the item location in the "item-only simulation"). Following this assumption, we generated 1000 simulations for each participant. Each simulation consisted of six simulated retrieved items, corresponding to all the six encoded landmark locations. For each encoded location, the simulated retrieved location is generated based on not only the participant's true error for this specific location, but also based on a probability assigned according to the distance from that simulated location to the reported center as follows (Fig. 2.7a). For any encoded location, the space where a simulated retrieved location can possibly be generated is a circle centering the encoded location with a radius of an error from participants' error data. We divided the circle into 200 angles. At each angle on that circle, we calculated the distance from that angle to the reported center and assigned a probability to that angle based on such distance:

$$P_i = \frac{d(simulated\ item_i,\ reported\ center) - (d(simulated\ item_i,\ reported\ center))}{\sum_{i=1}^{n} \ (d(simulated\ item_i,\ reported\ center) - \ max\ (d(simulated\ item_i,\ reported\ center)))}$$

where *i* corresponds to each angle and *n* corresponds to the total number of angles (200). The probability for any location to be retrieved would thus be inverse to the distance between the angle and the reported center. If a simulated location fell outside the

boundaries of the screen, the algorithm generated a new location. The bias value for each of the 1000 simulations was the average across each simulation's six retrieved locations. The gist simulated bias for each participant was the average across the 1000 simulations.

For both item-only simulations and gist simulations, when item memory error increases, the increased error will lead to a negative bias value despite no meaningful bias away from the center of the landmarks (Fig. 2.7c). Therefore, it is necessary to compare the data with the simulations.

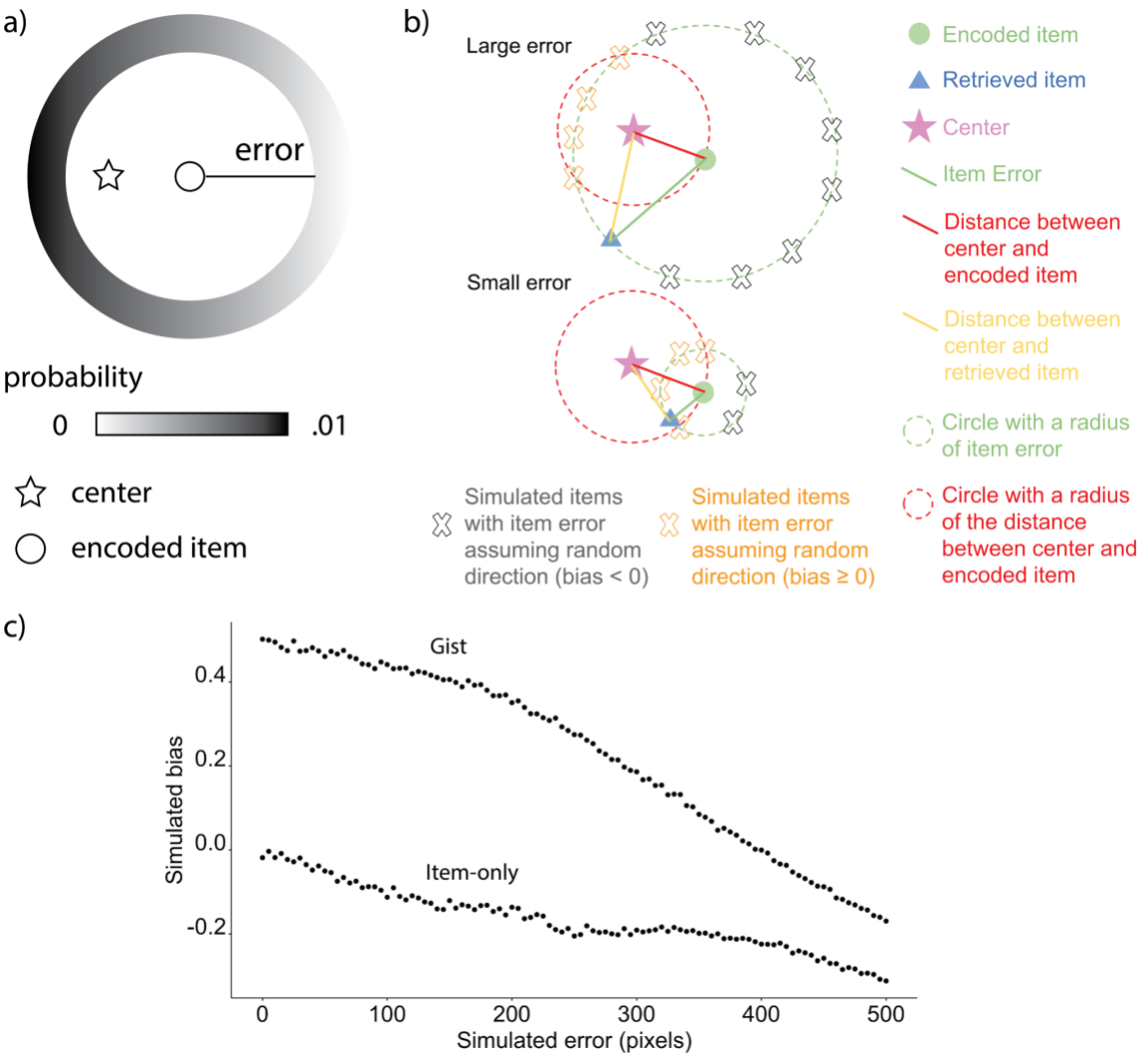


Fig. 2.7. Illustration of the simulations. a) An example of the probability of simulated locations to be generated for an encoded location given the same extent of error in gist simulation. b) Items with large errors are more likely to have a negative absolute bias by chance. First, the proportion of the arc with negative absolute bias in the circumference of simulated items is higher for items with large error. Because the distance between any points on the arc defined by the intersection between the green circle and red circle to the center will be shorter than the red distance, the points will all have an absolute bias value ≥ 0 (indicated by orange X marks), whereas the points outside of the arc will have a

negative bias (indicated by grey X marks). Second, even though the retrieved item (blue triangle in the lower figure) with small error and the retrieved item with large error (blue triangle in the upper figure) are biased in the same direction, the absolute bias for the retrieved item with the small error is positive whereas the other is negative, which demonstrates how a retrieved item with large error could cause negative bias without meaningfully being biased away from the center relative to its encoded location. c) Simulations based on the 6 encoded locations in Experiment 1 showed that random error of retrievals not assuming direction was negatively correlated with bias for both item-only simulation and gist simulation.

Statistics. To examine whether gist memory persisted when memory for items decayed over time, we conducted a 3 (group: 24-hour, 1-week, and 1-month) X 2 (memory type: item, center) aligned ranks transformation ANOVA, a nonparametric approach that allows for analyzing main effects and interaction (Kay, M., 2020), of the error change (Session 2 error values - Session 1 error values) and also two-tailed Mann-Whitney tests for between-group error change comparisons, because the data were not normally distributed as determined by a Shapiro-Wilk test. In order to examine whether there is a relation between item and gist memory, we used a linear model to evaluate the effects of estimated center error, delay group, and their interaction on reported center error. To interpret the effects of the overall effect of the delay group on the gist error, rather than the individual effects of each group, we reported the SSE rather than betas.

In order to examine whether the observed bias became more dissimilar to bias predicted by the item-only simulation over time, we conducted a 2 (session: Session 1 vs. Session 2)

x 3 (delay groups: 24 hours, 1 week, and 1 month) ANOVA in the difference between the observed bias and item-only simulated bias (bias data - item-only simulated data). As follow-up analyses, we used two-tailed paired t-tests to compare the difference in observed bias and item-only simulated bias between session 1 and session 2 for each delay group. In order to examine whether item retrievals were biased towards the center at any time points, we compared the observed bias against the item-only simulated bias for each group at each session.

In order to examine whether the observed bias became more similar to bias predicted by the gist simulated bias, we conducted the same ANOVA in the difference between the observed bias and gist simulated bias (gist simulated data - observed bias). As follow-up analyses, we used two-tailed paired t-tests to compare the difference in observed bias and gist simulated bias between session 1 and session 2 for each delay group. Note that due to the limited number of trials, we could not fit the most accurate parameter of gist influence and therefore the amount of gist influence under the gist simulation is arbitrary and likely not accurately reflecting the amount of gist influence in observed data. Therefore, unlike the analysis for item-only simulated bias, we did not predict that over time the observed bias would become indistinguishable from the gist simulated bias and tested whether observed bias and gist simulated bias significantly differed at all sessions for the delay groups. Reports were not corrected for multiple comparisons.

Follow-up Gist-based Bias Analyses.

Center of Encoded Locations as Bias Center: This analysis was exactly the same with the bias analysis in the prior section, except that this analysis used the center of encoded locations, instead of the center participants reported as bias center.

Weighted Center by Item Accuracy: This analysis was exactly the same with the bias analysis in the prior section, except that this analysis used a center weighted by item accuracy as bias center, instead of the center participants reported as bias center. The weight of each item was determined by (1- error of the item/error of all items)/(number of items - 1), such that items with higher accuracy would be weighted more in the computed center and also that the weight of all items summed up to 1.

Minkowski's Space: This analysis was exactly the same with the bias analysis in the prior section, except that all the distance was computed by $d(a,b) = \sqrt[1.5]{(a_1 - b_1)^{1.5} + (a_2 - b_2)^{1.5}}$. We selected 1.5 because Minkowski distance is defined by $\sqrt[g]{(a_1 - b_1)^{g} + (a_2 - b_2)^{g}}$ and g is typically selected between 1 and 2.

Swapped Items. In recalling the location for the landmarks, participants might "misbind" the label of a landmark and its location (e.g., indicate the location of the "restaurant" at the actual location for the "university" and vice versa). In order to test the potential influence of such errors on our results, we developed a criterion to identify pairs of items that were swapped, and we swapped them back to see if that changed the results. That is, for example, if (1) the retrieval for "restaurant" was closest to the encoded location for "university", (2) the retrieval for "university" was closest to the encoded location for "restaurant", (3) the retrievals were both within the range of both of the encoded locations (i.e. the distance

between encoded "restaurant" and "university" / 2) and, (4) there were no other retrievals in this range, we then swapped the retrieved university and restaurant responses and used the swapped results for the analyses described above. We found that swapping the items did not change any of the reported results.

Experiment 2

Participants and procedure. We recruited 77 members of the University of Pennsylvania community (18-30 years old; normal or corrected to normal vision) to participate in the experiment for monetary compensation. Sample size was based on prior behavioral memory studies[5]. All procedures were approved by University of Pennsylvania IRB (IRB #705915, Linguistic and Nonlinguistic Functions of Frontal Cortex). All 77 participants received training and testing during Session 1 and reported item and center memories again after 24 hours (Session 2). Sessions 1 and 2 were identical with Experiment 1, except that in Experiment 2 during Session 1, participants were trained to retrieve eight landmark locations, one of which was a spatial outlier (see Experiment 2 stimuli in Fig. 2.1), until their retrieval error for each landmark was fewer than 50 pixels (again, a distance less than ½ of the shortest distance between the pairs of encoded locations) in any direction. In addition, in Session 1 of Experiment 2, to streamline the session, item memory was derived from the item memory test in the last round of evaluation during training (Phase 3), which was immediately followed by the gist memory test (see Experiment 2 procedure

[5] Because this was a new experiment, we were unable to identify an effect size from a past study that was appropriate for a power analysis. Therefore, we tried to collect a sample equivalent to what is commonly collected in behavioral memory studies (e.g., Schapiro et al., 2017) and continued to recruit participants across two semesters.

in Fig. 2.1). The time for Session 1 was approximately 25 minutes, which was longer than that for Experiment 1 because in Experiment 2, participants learned more locations and the training criterion was harder (50 pixels, as opposed to 80 pixels in Experiment 1).

After 32 to 57 days, 50 participants returned for Session 3 by email invitation. Session 3 was identical to Session 2 (i.e., participants reported their memory for the center and then each item). The time for Session 2 and 3 was approximately 10 minutes. Of the 50 participants who returned for the third session, 1 participant was excluded because their individual and gist performance for at least one session was lower than 3 *SD* below average. We did not exclude participants whose reported gist memory error was larger than the distance between the screen center and the true center at Session 1, as in Experiment 1, because in Experiment 2, a large gist error could be a meaningful result that reflects the overweighting of the outlier in reporting gist. We excluded 6 participants who placed the outlier where the majority of items were, which means the error of the outlier was larger than 573 pixels (i.e., the distance between the center of screen and outlier encoded location). The reason we excluded these participants was that in Experiment 2, if participants swapped the outlier with one of the other items, or simply put the outlier among the other items, this outlier swap would strongly inflate the bias value towards the global reported center, which does not necessarily reflect a true bias towards the center. Our reported results thus include 43 participants (age: $M = 21.5$, $SD = 2.2$, gender: 75% female)

Error measurement. All error measures were calculated as in Experiment 1, except that the chance performance for individual items based on the center of the screen was 386 pixels (determined by the average Euclidean distance between the center of the laptop screen and

each encoded item location), chance performance for items based on center of encoded locations was 262 pixels (determined by the average Euclidean distance between the center of the encoded item locations and each encoded item location), and the chance performance for gist memory was 223 pixels (determined by the average Euclidean distance between the center of the laptop screen and the true center of all encoded item locations).

Bias Measurement. We calculated bias as in Experiment 1, except that we additionally computed a local gist bias, which was a bias index using the local center (i.e., the center of the seven encoded locations excluding the outlier) as the bias center.

Outlier weight estimation. In order to estimate the weight of the outlier on the reported gist memory, we developed a weight model as follows: For each participant at each session, a series of weights were applied to the encoded outlier location. The range of the weight of the outlier was from 0 to 1, with a stepwise increment of 0.0125. The weight for each of the other encoded items was assumed to be the same and would thus be (1 - outlier weight)/number of items that were not the outlier, ranging from 0 to 0.125. Based on these weights, 81 simulated centers were computed: when the outlier weight was 0, the simulated center would be a perfect local center ignoring the outlier; when the outlier weight was 0.125, the simulated center would be a perfect global center of all items, since there were 8 items in total; when the outlier weight was 1, the simulated center would be the outlier itself.

For each participant at each session, the Euclidean distance between each simulated center and reported center was computed, resulting in 81 distances. We used the weight

that resulted in the smallest distance as the estimated weight of the outlier for that participant at that session.

Statistics. As in Experiment 1, we conducted a 2 (delay: short retention of 1 day or long retention of 1-2 months) x 2 (memory type: item, center) aligned ranks transformation ANOVA with repeated measures of error change (short, defined by Session 2 error values - Session 1 error values, or long, defined by Session 3 error values - Session 1 error values) and pairwise Wilcoxon signed rank tests for error comparisons, since change in error was not normal as determined by a Shapiro-Wilk test.

As in Experiment 1, in order to examine whether there was a relation between item and gist memory, we used a linear mixed effects model on reported center error with fixed effects of delay (24 hours and 1 to 2 months), estimated center error, and their interaction, as well as a random effect of participant.

As in Experiment 1, in order to examine whether the item retrievals were increasingly biased towards the reported center over time, we conducted a one-way repeated measures ANOVA in the difference between observed bias and item-only simulated bias across sessions (after training, 24 hours, and 1 month). We compared the bias values against the item-only simulated bias controlling for error against 0 at each session by paired t-tests to examine whether item retrievals were significantly biased towards the reported center. We conducted the analogous ANOVA analysis in the difference between observed bias and gist simulated bias using reported center as bias center. As explained in the results, we then conducted the same statistical analyses using the local center as the bias center and then

follow-up paired t-tests between sessions. Reports were not multiple comparisons corrected.

In order to examine whether the outlier was weighted more, the same, or less compared to other items, we compared the outlier weight values against the weight assuming all items to be equal (i.e., ⅛ = 0.125) with t-tests. In order to examine whether the weight of the outlier in gist memory changed over time, we used a paired t-test comparing the outlier weight change after a short retention interval (Session 2 outlier weight values - Session 1 outlier weight values) against the outlier weight change after a long retention interval (Session 3 outlier weight values - Session 1 outlier weight values).

Data availability

All data will be available at

https://osf.io/jxme8/?view_only=049dcb1efaf44c3098040ba027f88115

Code availability

All scripts used to analyze the data will be available at

https://osf.io/jxme8/?view_only=049dcb1efaf44c3098040ba027f88115

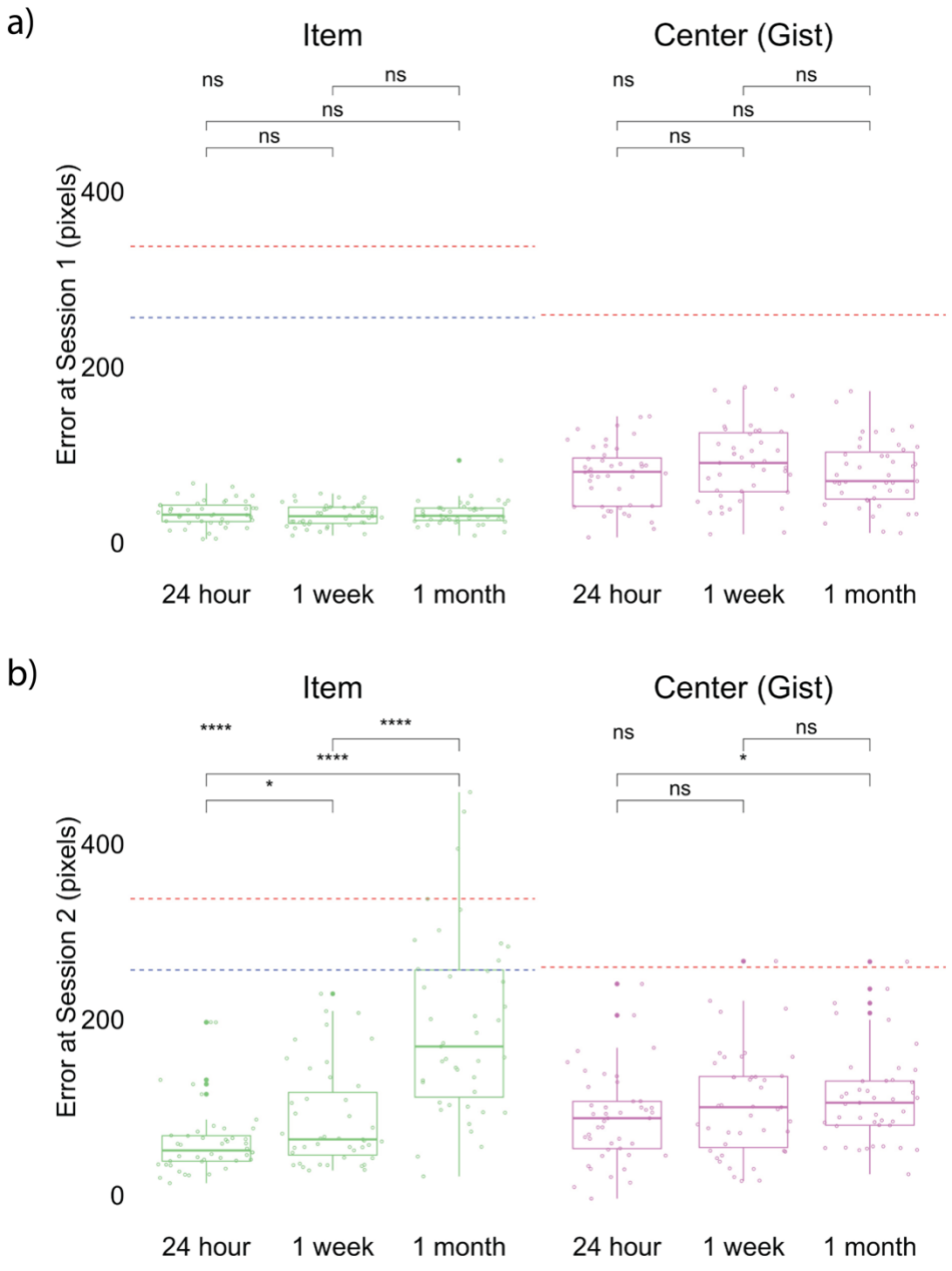


Figure 2.8. Error at each session in Experiment 1. Experiment 1 error in item and gist (center) memory at Session 1 (a) and at Session 2 (b). * indicates $p < .05$, **** indicates $p < .0001$, and ns indicates $p > .05$ by t-tests between groups and ANOVA (top left). Red dashed lines indicate chance performance for item (defined as the average of distance between encoded item locations and center of the screen) and gist memory (defined as the

distance between the center of encoded locations and center of the screen) based on center of the screen. Blue dashed lines indicate chance performance for item memory, based on the center of encoded locations. This corresponds to what participants' performance would be if they only remembered the center and just clicked the center when asked to recall an item. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers.

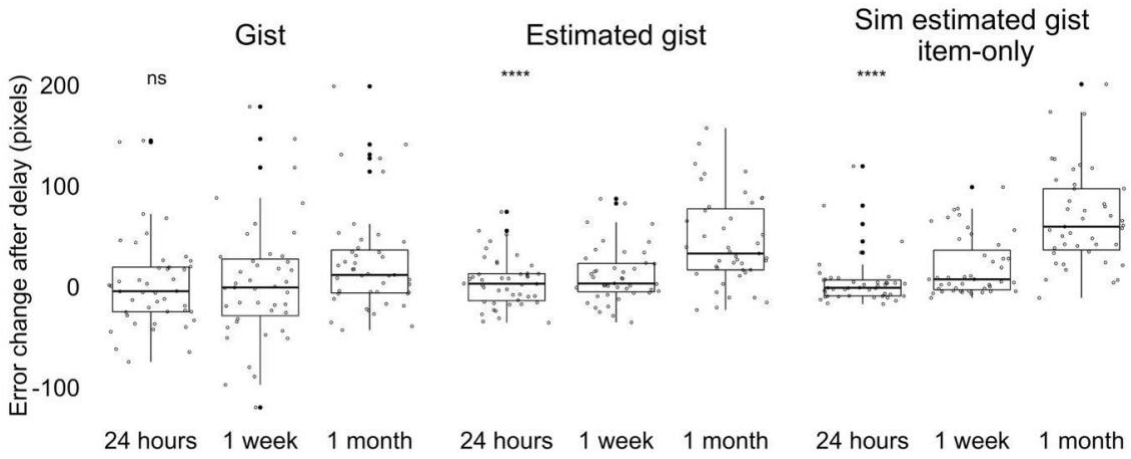


Figure 2.9. Experiment 1 error change between Session 1 and 2 in gist (center) memory, estimated gist, and simulated estimated gist based on a simple item-only simulation, assuming that the magnitude of error for each item memory would remain the same but the direction of error would not be systematically influenced by the gist[6]. Aligned rank transformed ANOVA analysis with three gist memory error type revealed a main effect of delay, $F(2, 381) = 42.93$, $p < .001$, memory type, $F(2, 381) = 15.17$, $p < .001$, and an interaction between delay and memory type, $F(4, 381) = 3.83$, $p < .01$, suggesting that the error increase over time is not the same for these gist memory types. For reported gist error (Gr) and simulated estimated gist error (sGe), we found a significant interaction between delay group and gist memory error type, $F(2, 254) = 6.68$, $p = .001$. For

[6] We generated 1000 simulations for each participant. Each simulation consisted of all simulated retrieved items, corresponding to all the landmark locations. For each location, we randomly generated a retrieved location based on the participant's true error for this specific location, allowing angle to vary randomly across the simulations. Then, we computed the center for these locations to get the simulated estimated gist for each simulated participant. The error for such simulated estimated gist was the Euclidean distance between the true center and the simulated gist. The simulated estimated gist error for each real participant was the average value of simulated estimated gist error for their corresponding 1000 simulated participants.

estimated gist error (Ge) and simulated estimated gist error, we also found a significant interaction between delay group and gist memory error type, $F(2, 254) = 3.28$, $p = .039$. Gist error and estimated gist error both increased less over time compared to the simulated estimated gist error under this simple item-only simulation over time, suggesting that participants' data are not compatible with this simple item-only simulation. We did not find a significant interaction between Gr and Ge across time, $F(2, 254) = 1.18$, $p = .31$, consistent with the idea that Ge was calculated from item memories influenced by the center after delay. **** indicates $p < .0001$ and ns indicates $p > .05$ by ANOVA (top left). The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers.

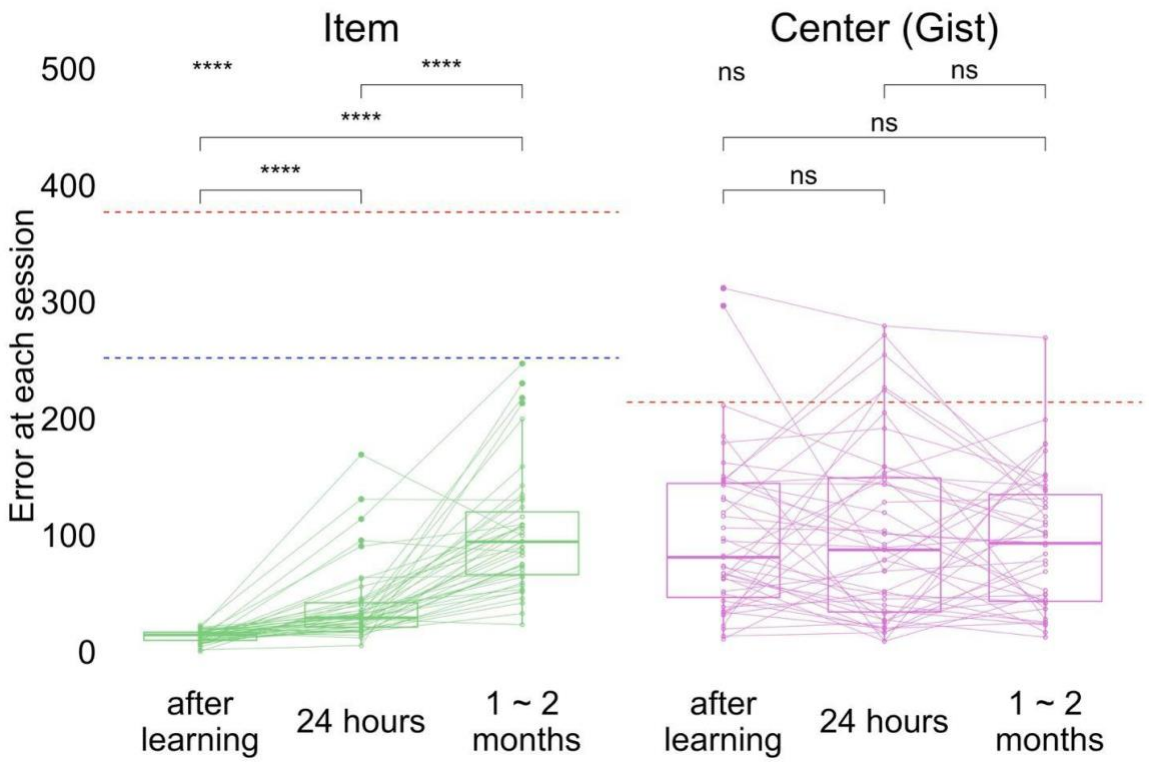


Figure 2.10. Experiment 2 error in item and gist (center) memory at each session. Red dashed lines indicate chance performance for item (defined as the average of distance between encoded item locations and center of the screen) and gist memory (defined as the distance between the center of encoded locations and center of the screen) based on center of the screen. Blue dashed lines indicate chance performance for item memory, based on the center of encoded locations. **** indicates $p < .0001$ by paired t-tests and ANOVA (top left). The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers. Dots and lines indicate participants.

Chapter 3: Item distinctiveness influences gist memory formation

Abstract

How do humans synthesize and aggregate across individual experiences to form summary statistics? It remains unclear how the properties of the items, such as reward motivation, influence how these items contribute to the gist. In two experiments, 500 participants encoded spatial locations that are emphasized with different sources of distinctiveness (reward motivation, increased attention by a distinctive color, and repeated exposure). After this training, participants reported their item memory (memory of these locations) and their gist memory (memory of the center) for these locations. Experiment 1 found that reward motivation increased the accuracy of the distinctive item without changing the overall item and gist memory. Experiment 2 discovered that increased attention by color and repeated exposure distorted gist memory in different ways, while reward motivation preserved the gist memory. Our results demonstrate that different sources of item distinctiveness influence gist memory formation and suggest a protective role of reward motivation on gist memory.

Introduction

Human minds have the extraordinary ability to extract summary statistics from observing and experiencing individual events and instances. For example, you probably have seen many pies in your life, but when you are asked to think about an average pie, something round and brownish may come to your mind. This ability to extract the summary statistics, to "get the gist of it", is important to our life. Following the example

of pie, in order to decide whether you would like to eat a pie tonight, you probably need to have a memory of the average pie. Given the importance of summary statistics, little is known about how human minds synthesize and summarize these individual experiences to form this memory of average. For instance, if the pies your grandmother makes always give you sweetness and warmness, would the memory of these pies become more memorable, and thus shape your memory of the average pie more than the other pies?

Much research has demonstrated that humans are excellent at extracting and retaining the "gist memory", the memory of summary statistics across individual instances, over time (Graves et al., 2020; Lutz et al., 2017; Posner & Keele, 1970; Richards et al., 2014; Zeng et al., 2017). However, not much research has examined how the individual items contribute to the gist, and particularly the factors that may influence such contributions in long-term memory.

Ensemble perception work has provided evidence that not all items contribute to the gist equally in working memory (Alvarez, 2011; Whitney & Leib, 2018). This line of research usually operationalizes item memories as item properties on a continuous dimension such as the size of individual circles and gist memories as the average of the property such as the average size of the group of circles (Brady & Alvarez, 2011). They show that distinctive items, such as items that received more attention or items that are deviant from other items, disproportionately influenced the gist. For example, emotional faces with more attention contribute more to the average emotional faces that participants recalled (Ying, 2022). Expressions that are deviant from other expressions are discarded in memory of the average expressions (Haberman & Whitney, 2010).

A few long-term memory studies have also suggested a disproportionate influence of distinctive items on the gist memory. “Outlier” items, which are spatial locations that are far from other locations, greatly distorted the memory of the summary statistics across these spatial locations for both rodents and humans (Richards et al., 2014; Zeng et al., 2021). These studies offered paradigms that can measure the contribution of items to the gist in long-term memory. Through operationalizing item memory as the memory of spatial locations on a computer screen and gist memory as the memory of the center of these spatial locations, the values of the weight of these items in the gist can be estimated or calculated from participants’ retrieved memory (Zeng et al., 2021), similar to working memory literature (Haberman & Whitney, 2010). These paradigms provide a start for us to understand the factors that influence how each item contributes to the gist in the long-term memory.

It will be particularly meaningful to investigate reward motivation as a source of distinctiveness. Work in motivational learning has demonstrated that emphasizing items with reward motivation during learning on items improves the memory of these items (Cowan et al., 2021; Shigemune et al, 2014). Nevertheless, less is understood about how this reward motivation influences the extraction of gist memory across these items. Reward motivation on items may change the gist memory formation. Neuroimaging work shows that reward during learning facilitated the coupling between the hippocampus and ventromedial prefrontal cortex (vmPFC) (Murty et al., 2016). Interestingly, this coupling between hippocampus and vmPFC is associated with gist memory extraction (Tse et al., 2007; Zeithamova et al., 2012).

The recruitment of a similar neural network may lead to an influence of reward motivation on the formation of gist memory at the behavioral level. It is unclear how. Reward motivation of items may improve the accuracy of gist memory by facilitating the neural network of gist extraction. On the other hand, the reward may impair the accuracy of gist memory by utilizing the same resources of the neural network.

Moreover, reward motivation may influence the gist by altering the contribution of the emphasized item to the gist memory. It is possible that reward will increase the weight of items in the gist by increasing the attention or the accuracy of these items (Yang, 2022). Alternatively, reward motivation may decrease the weight of the emphasized item because reward may make the emphasized item different from the other items, and thus separate that emphasized item from being incorporated into the gist (Haberman & Whitney, 2010). A third possibility is that items emphasized with reward may contribute to the gist to the same level as other items because reward protects the gist memory from being distorted by the increased distinctiveness of some items (Clewett & Murty, 2019).

The current study intended to disentangle these possibilities. In Experiment 1, we trained participants to learn two sets of spatial locations. One set of locations contained an item that was associated with monetary reward during training (the reward condition) and the other set of locations did not (the neutral condition). We contrasted participants' memory of the spatial locations as well as their memory of the center of these locations between these two conditions in order to understand the influence of reward motivation on gist memory extraction. To explore the influence of reward motivation on the contribution of the items to the gist, we computed the weight values of the reward item in the gist memory.

In Experiment 2, we aimed to separate the influence of the reward from that of increased attention and accuracy. We contrasted the item memory, gist memory, and the weight values of the distinctive item in the gist in the reward condition with not only these memories in the neutral condition, but also these memories in an attention condition where an item was emphasized with increased attention, and in a frequency condition where an item was emphasized with repeated exposure. Taken together, these experiments provide insights into how different sources of distinctiveness, especially reward motivation, influence the extraction of summary statistics from individual experiences.

Results

Experiment 1

66 participants completed the experiment in the lab **(**age: M = 19.5, SD = 6.7, gender: 79% females). In Experiment 1, item memories were operationalized as two categories of locations on a laptop screen, six 'landmarks' and six 'animals' (i.e. dots associated with unique names of landmarks and animals) (Figure 3.1b). Participants learned these categories of locations through training (Figure 3.1; see Methods for details. The reward and neutral conditions only differed in the training of a particular item: in the reward condition, participants were instructed that remembering one specific item during the training would work towards earning monetary bonuses, and the item name would be marked with signs of "$" during training, whereas in the neutral condition, all the items were learned in the same way with no monetary bonuses.

After the training, participants were instructed that all the parts associated with monetary bonuses had ended, but they were still encouraged to do their best in the task. They were only able to proceed after they solved 10 arithmetic questions, which was intended to minimize the influence of working memory. Finally, they were tested on item and gist memory for each condition. During the gist memory test, participants reported their memory of the center (average location) of a category. During the item memory test, they recalled each item location in random order with no feedback. The order of reward and neutral conditions corresponded to the order of them during the training phase. Through the experiments, we collected participants' memory of the items and their memory of the gist.

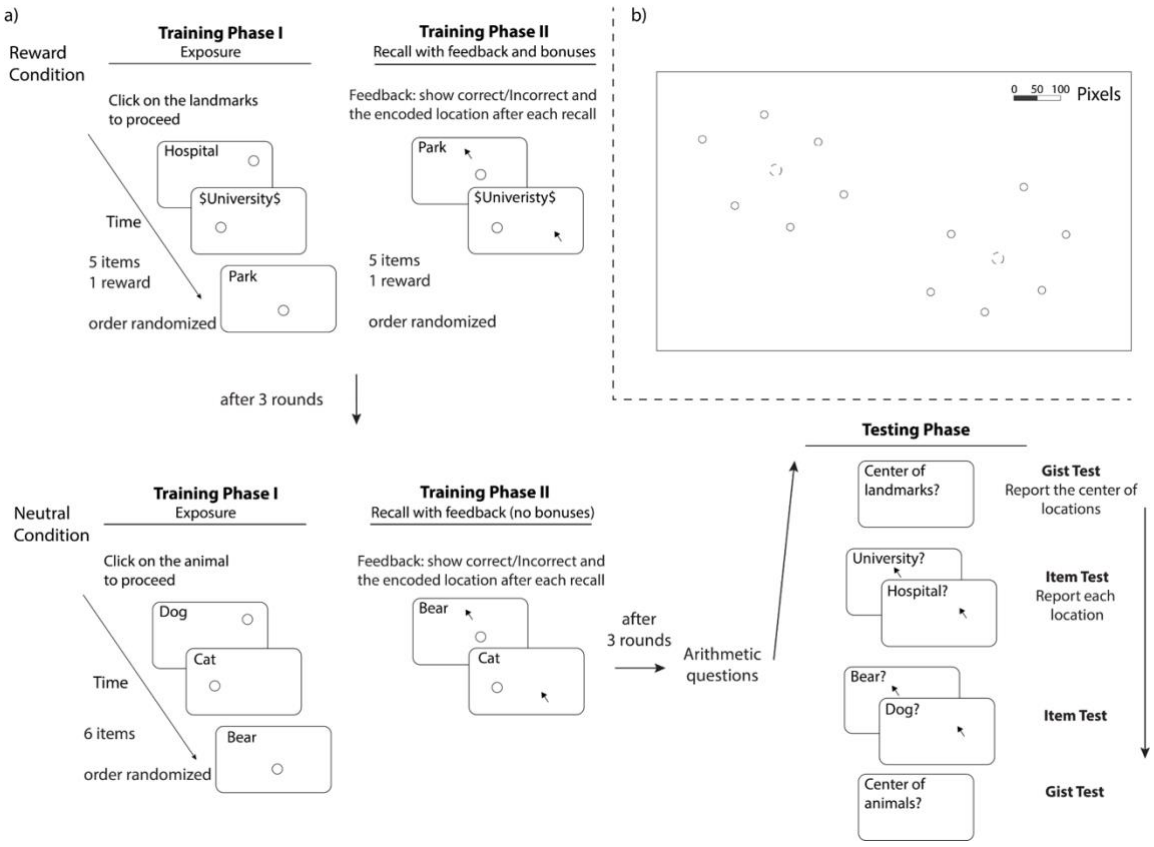


Fig. 3.1. Procedure and stimuli for Experiment 1. (a) An example of the procedure. The order of reward condition and neutral condition and the order of item and gist tests within each condition were both randomized across participants. (b) An illustration of the location of the stimuli (drawn to scale). The circles in solid lines indicate the locations participants learned. The circle in dashed lines indicates the center of each category/condition of locations. The assignment between category and the cluster of locations and the assignment between category and condition are randomized across participants.

The reward and the neutral condition did not differ in overall gist and item error, but the reward item had higher accuracy

Our analyses did not find any significant differences between the reward condition and the neutral condition in their overall item memory error (Figure 3.2a; Figure 3.2c) and gist memory error (Figure 3.2a; Figure 3.2b), Wilcoxon signed-rank tests: *ps* > 0.17. The error of the reward item in the reward condition was significantly lower compared to that of the neutral item in the reward condition, Wilcoxon signed-rank test: $Z = 5.47$, $p < 0.0001$, and also that of the items in the neutral condition, Wilcoxon signed-rank test: $Z = 4.88$, $p < 0.0001$ (Figure 3.2d). The difference between the neutral items in the reward and the neutral condition was not significant, Wilcoxon signed-rank test: $Z = 1.91$, $p = 0.06$. These results suggest that the reward on an item successfully increased the accuracy of that particular item, but did not change the overall accuracy of the item and gist memory.

To demonstrate the relation between item and gist memory that was detected in prior studies (Zeng et al., 2021), we used a linear model to evaluate the effects of item memory error and the condition on reported center error. We found that item memory error significantly predicted gist memory error, $SSE = 4502.6$, $F(1, 108) = 6.75$, $p = 0.01$. The interaction between the item memory error and condition on reported center error was not significant $SSE = 929.4$, $F(1, 108) = 1.39$, $p = 0.24$. These results indicate a positive relationship between the item and gist memory for both the reward and the neutral condition.

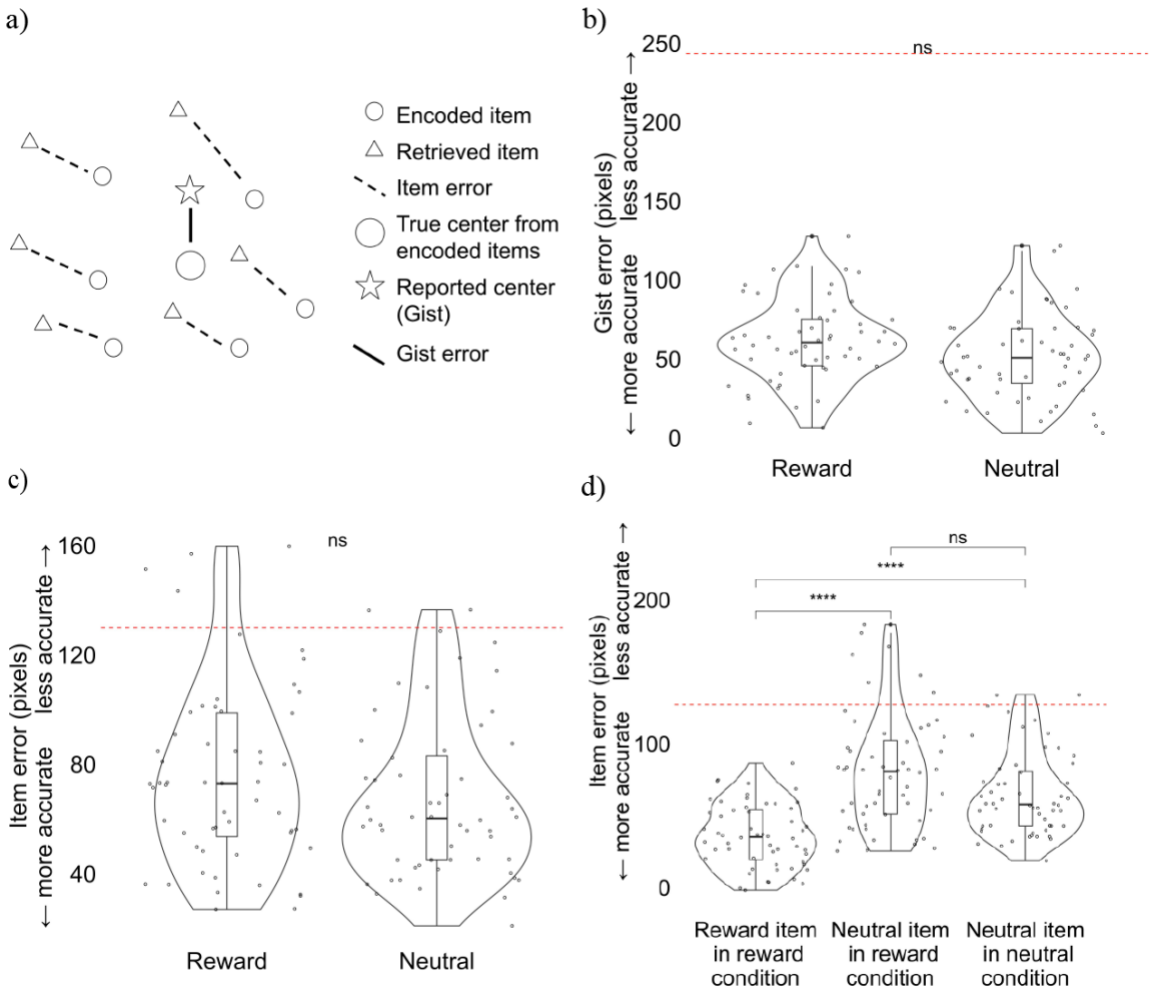


Fig. 3.2. The error of participants' item and gist memory. (a) Error measurements. (b) Participants' gist memory error by condition. Red dashed lines indicate chance performance for gist memory (defined as the distance between the center of encoded locations and the center of the screen). (c) Participants' item memory error by condition. (d) Participants' item memory by the type of the item. *** indicates $p < 0.001$, **** indicates $p < 0.0001$, and ns indicates $p > 0.05$ by paired Wilcoxon signed-rank tests between conditions (b and c) and between item type (d). For c and d, red dashed lines indicate chance performance for item memory (defined as the average distance between encoded item locations and the center of the screen). This corresponds to what participants' performance would be if they clicked the screen center whenever asked to recall an item. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers.

The reward item did not contribute more or less to gist memory

In order to examine how the reward on an item will influence its contribution to the gist memory, we computed the weight values of the reward items. This computation was adapted from weighted models in prior work on working memory and long-term memory (Haberman and Whitney, 2010; Zeng et al., 2021; see Methods for details). The weight of the reward item was marginally lower than 0.167 (i.e., ⅙, the level assuming equal weights across all items), Wilcoxon signed-rank tests: $Z = 1.83$, $p = 0.07$, (Figure 3.3). It can be concluded that the reward item did not contribute more to the gist memory compared to other items.

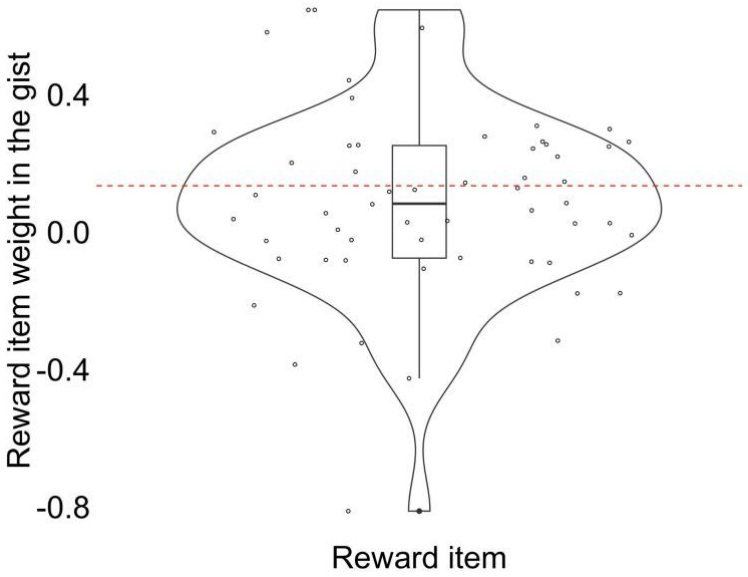


Fig. 3.3. The weights of the reward item in the gist memory. The red dashed line indicates the chance level when the weight of the reward item when equal to the weight of other items in the gist memory. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers). Greater values indicate higher weights of the reward item in the gist memory.

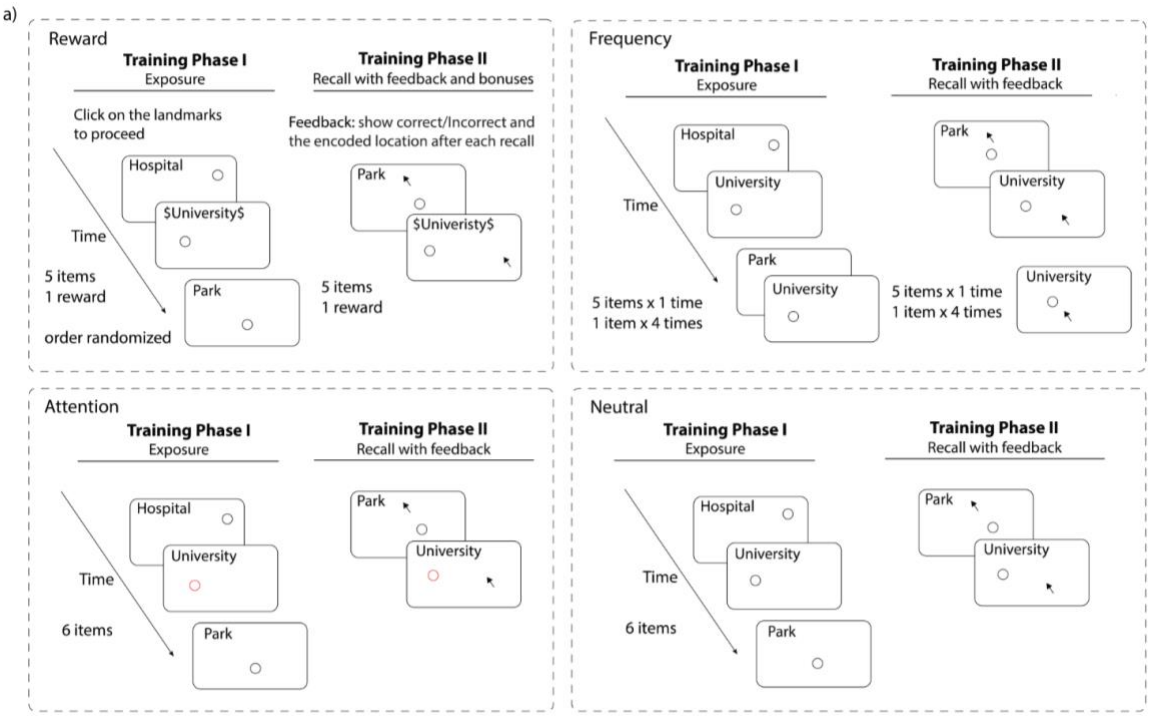
In order to understand whether the accuracy of the reward item influences the weights, we computed the correlation between the weight and the accuracy of the reward item. We did not find a significant correlation (p = 0.31). There is no evidence from Experiment 1's data for an influence of item accuracy on its weight to the gist memory. Taken together, results from Experiment 1 suggest that the reward motivation on an item increased the accuracy of that item, but did not influence the overall item and gist memory.

Experiment 2

To further understand how the distinctiveness of items influences the gist memory, we conducted Experiment 2. The training procedure and stimuli of Experiment 2 (Figure 3.4; see Methods for more details) were similar to those of Experiment 1 (Figure 3.1). However, the two experiments differed in three major aspects (see Methods for more details). First, in order to further understand the role of reward on the gist memory, we conducted two additional distinctiveness conditions to contrast with the reward: the attention condition and the frequency condition (Figure 3.4), to control for the influence of attention and accuracy. The four conditions were the same except for one critical difference: how one of the six items is emphasized. The reward condition emphasized one item with a monetary bonus of $3 that will double the payment; the frequency condition emphasized one item with repeated training; the attention condition emphasized one item with the color red, utilizing the Von Restorff effect to increase attention on this item (Schmidt & Schmidt, 2017); the neutral condition did not emphasize any items, but had one foil item which matched exactly with the emphasized items in other controls. We assigned each participant a stimuli ID, through which we can identify the stimuli they get and match across conditions. Second, Experiment 2 had an across-participants design, where each participant only experienced one condition and learned one set of six locations. This was to avoid any influence between the two clusters of locations. Third, in Experiment 1, it was unclear whether there was any noise caused by the stimuli that would impact the weight of the emphasized item in any particular way. In order to control for this possible noise and enable direct comparisons across all the conditions, Experiment 2 deployed a “yolked” design, where each participant in each of the

conditions received exactly the same stimuli as what corresponding participants in other conditions would receive, including the location-name mapping, the presentation order of the locations in each round, the order of the item and the gist test, which item was selected to be emphasized, etc. After the training, participants reported their gist and item memory, with the test order varied across participants, similar to Experiment 1.

Our reported results thus included 422 participants who completed the task on Amazon Mechanical Turk. Each participant was randomly assigned to one of the four conditions: reward, frequency, neutral, and attention (Figure 3.4). Our reported results include 422 participants, with 112 participants in the reward condition, 102 participants in the repeat condition, 108 participants in the attention condition, and 100 participants in the neutral condition.



b)

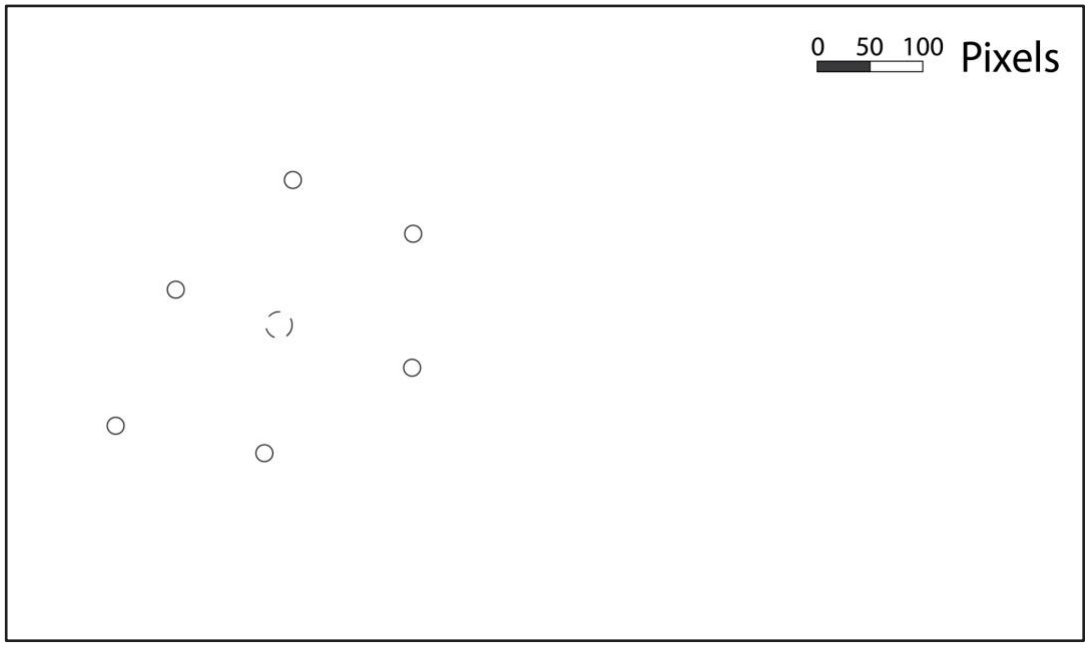


Fig. 3.4. Procedure and stimuli for Experiment 2. (a) Training procedure for the four conditions. The arithmetic questions and the test phase are similar to Experiment 1 and thus are omitted here. (b) An illustration of the location of the stimuli (drawn to scale). The circles in solid lines indicate the locations participants learned. The circle in dashed lines indicates the center of each category/condition of locations.

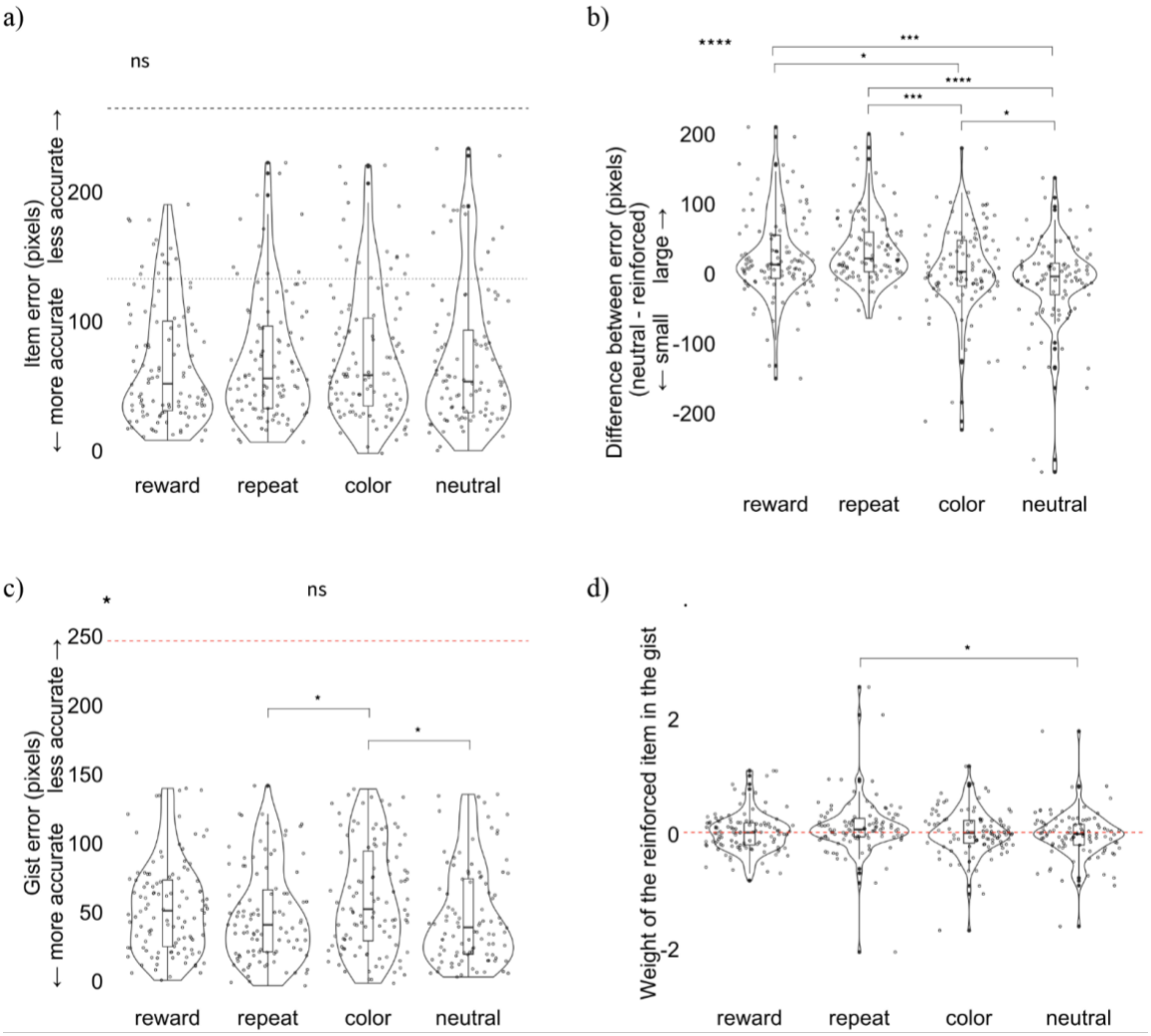


Fig. 3.5. Errors and weight values by condition. (a) Participants' item memory error by condition. Dashed lines indicate chance performance defined as the average distance between each encoded item location and the center of the screen. Dotted lines indicate the chance performance defined as the average distance between each encoded item location and the center of these encoded locations. (b) The difference between the error in the neutral items and the emphasized item. A larger difference indicates higher accuracy for the emphasized items compared to the neutral items. (c) Participants' gist memory error by condition. Red dashed lines indicate chance performance for gist memory (defined as

the distance between the center of encoded locations and the center of the screen). (d) The weights of the reward item in the gist memory by conditions. The red dashed line indicates the chance level where the weight of the reward item is the same as that of other items in the gist memory. Greater values indicate higher weights of the reward item in the gist memory. For all figures, *** indicates $p < 0.001$, **** indicates $p < 0.0001$, . indicates $p = 0.08$, and ns indicates $p > 0.05$ for a main effect of condition between each pair as well as all conditions together (top left) by linear mixed-effects models controlling for stimuli. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers.

Sources of distinctiveness did not change overall item accuracy but increased the accuracy of the emphasized item at different levels

In order to understand the influence of distinctiveness on the item memory (see Methods for details of how it is calculated), we conducted a linear mixed-effects model on overall item memory error (Figure 3.5a) with fixed effects of conditions (reward, attention, frequency, and neutral) and a random effect of stimuli ID to account for the noise from the stimuli. The main effect of condition, $SSE = 1532$, $F(3, 418) = 0.21$, $p = 0.89$, which suggests that the four conditions matched in their overall item memory error.

However, all the distinctive items (reward, attention, and frequency) had higher accuracy compared to the neutral items, which was shown by the following statistical analyses. We computed the difference between the neutral item error and the distinctive item error for each condition. The difference scores for the three distinctiveness conditions were all higher than 0 (Wilcoxon signed-rank tests: ps < .001), but the

difference score for the neutral conditions was not significantly different from 0 (p = 0.17). These results suggested that all the distinctiveness manipulation successfully increased the accuracy of the distinctive items.

In addition, we conducted a linear mixed-effects model on this difference score with fixed effects of conditions (reward, attention, frequency, and neutral) with a random effect of stimuli ID to account for the noise from the stimuli. We found a significant effect of condition, SSE = 123954, $F(3, 418) = 12.12$, $p < 0.001$ (Figure 3.5b). We then conducted the same fixed effects models on the difference scores for pairwise comparisons between conditions to further understand the source of the variation. In particular, the reward and repeat condition did not significantly differ in the difference scores, SSE = 6003.2, $F(1, 160.34) = 2.12$, $p = 0.15$, but all the other pairwise comparisons between conditions were significant, ps < 0.04 (Figure 3.5b). All the distinctiveness conditions had higher difference scores compared to the neutral conditions (ps < 0.04). These results suggested that the different types of distinctiveness did not influence the overall item memory, but increased the distinctive item accuracy to various extent.

Attention, but not reward and frequency, on the distinctive item impaired the accuracy of overall gist memory

In order to understand how the different sources of distinctiveness influence the gist memory, we conducted a linear mixed-effects model on the reported center error with fixed effects of conditions (reward, attention, frequency, and neutral) with a random effect of stimuli ID to account for the noise from the stimuli. We found a significant

effect of condition, SSE = 10779, $F(3, 368.79) = 2.93$, $p = 0.03$, which suggests that the four conditions differed in their gist memory error, despite the similarity of overall item memory across these conditions. In order to further understand the source of the main effect and separate the influence of reward, we compared pairs of these conditions separately with similar linear mixed-effects models. We found a significant difference between attention and neutral condition, SSE = 8528.5, $F(1, 206) = 6.17$, $p = 0.01$, and a significant difference between attention and frequency condition, SSE = 7889.8, $F(1, 208) = 5.87$, $p = 0.02$. These results suggested that only attention to particular items changed the overall gist memory error.

Frequency, but not reward and attention, increased the weights of the distinctive item in gist memory

To understand how sources of the distinctiveness of items influence the items' contribution to the gist memory, we computed the weights of the distinctive items in the gist memory for all four conditions (Figure 3.5d; see Methods for details) and compared these weights to chance level, ⅙, which was defined by what the weight of the distinctive item would be if it were the same with all the other neutral items. Only the weight in the frequency condition was significantly higher than the chance level ($Z = 2.57$, $p = 0.01$). This result suggests that the repeated exposure of a particular item during learning increased its contribution to the gist memory. Despite similar increased accuracy in the distinctive items, the rewarded item and the attention item both did not contribute to the gist memory as the repeatedly exposed item did (ps > 0.60).

In order to account for the possible noise from the stimuli, we conducted a linear mixed-effects model on weights of the distinctive item with fixed effects of conditions (reward, attention, frequency, and neutral) with a random effect of stimuli ID. The main effect of the condition was only marginally significant, $SSE = 1.11$, $F(3, 418) = 2.21$, $p = 0.09$. However, the weight of the distinctive item in the frequency condition was still higher than that of the neutral condition, $SSE = 1.04$, $F(1, 200) = 5.24$, $p = 0.02$. These results suggested that despite similar accuracy of the distinctive item, repeated exposure, but not reward, increased the weight of the distinctive item in the gist memory.

Taken together, results from Experiment 2 showed that increased attention, repeated exposure, and reward on an item all do not influence overall item memory accuracy, but strengthen the accuracy of the distinctive items. Despite having similar accuracy on item memory with reward, attention on an item impaired the gist memory accuracy compared to the neutral condition. Repeated exposure, as opposed to reward, increased the weight of the distinctive item in the gist memory, despite similar accuracy of this distinctive item between the conditions. Results suggest that the reward may serve a protective role to prevent the gist memory from being distorted.

Discussion

With two behavioral experiments with more than 500 participants, we examined how humans extract generalities across individual items and how sources of the distinctiveness of items influence the extraction of gist memory. We emphsized an item with different sources of distinctiveness: reward, attention, and repeated exposure. We tested the influence of the distinctiveness on item memory, gist memory, and how the

items contribute to the gist. Across the two experiments, we found that the distinctiveness of the items did not influence the overall item accuracy, although they increased the accuracy of the emphasized items. We found that both increased attention and exposure distorted the gist memory by either impairing the accuracy of the gist or shifting the weights of the emphasized item in the gist memory. In contrast, we did not find such effects for the reward condition. Our results highlight the different influences of the distinctive items on the extraction of gist memory and suggest a possible protective role of reward on gist memory.

We systematically manipulated memory of items with different sources of distinctiveness: reward, attention, and repeated exposure. We found that repeated exposure led to an increased weight of items in the gist. Although the item with reward and the item with repeated exposure shared the same level of accuracy, the reward item did not produce a higher weight on the items. If an increased accuracy of a emphasized item would lead to an increased weight for the item, then the existence of a reward may prevent this distortion by balancing and integrating the items in forming the gist memory. It may be argued that the attention also increased the accuracy of the emphasized items without increasing the weight of the items. However, the manipulation of attention did not increase the accuracy of the emphasized item to the same level as our reward and frequency condition (Figure 3.5b; Figure 3.9), and thus cannot provide a direct comparison for the weight. Future research can explore different ways of attention manipulation that reach the same level of accuracy as the reward and frequency conditions. On the other hand, although our attention manipulation resulted in the same level of overall item memory accuracy as other conditions, it worsened the gist memory

accuracy. The attention manipulation with a different color may disrupt the integration of items into the gist memory by separating them from other items or by allocating too much resource to an item and sacrificing the global processing of the items. This result adds to the limited evidence that increased attention can sometimes impair memory (Fu et al., 2021), and provide evidence for the interaction between memory and attention (Chun & Turk-Browne, 2007).

Our results show that different sources of distinctiveness on items influence the formation of gist memory differently and imply a protective role of reward on gist memory extraction which prevents the gist memory from being distorted. It is unclear whether the results are specific to the nature of our stimuli and procedure. For instance, the increased weight of the item of repeated exposure may be because repeated exposure increased the temporal frequency of the emphasized item in the visual system and thus registered this item into the gist uniquely. More work can be done to manipulate the accuracy by reinforcing the items with other sources of distinctiveness and disentangling the difference between frequency and accuracy. As another example, our experiments manipulated attention with the distinctiveness of color, and it is possible that other sources of perceptual and conceptual distinctiveness (e.g., the emphasized item can be told to be "dangerous") (Schmidt & Schmidt, 2017). Furthermore, it would be interesting to contrast our results with another source of motivation - threat. It was shown that reward and threat motivations recruit separate neural networks (Murty et al., 2016), and it is unclear how this difference will manifest in gist memory formation. Further research can be done to broaden the generalizability of the finding by investigating other sources of distinctiveness. We believe our findings provide a start for a promising research

endeavor to understand how different sources of distinctiveness will influence the integration of items to form the memory of generalities.

Our work established that not all items weigh equally in long-term memory. Prior literature has shown an enhancement of item memory associated with reward motivation after a delay (Patil et al., 2017). Relating this line of work to our paradigms and providing evidence on how the weights of items may change in memory consolidation after sleep and delay will contribute to the theories of memory consolidation. Our findings of differential weights in long-term memory also have potential implications for other fields of psychology, such as social stereotypes. We demonstrated that repeated exposure to a particular item increased its weight in the gist memory in the lab. Will this finding be able to be generalized to stereotype formation? For example, will repeated exposure to an instance in a particular social category (e.g., gender and race) on social media shape the stereotype of the social category more? Integration of theories and methods of the cognitive mechanism of differential weighing into social categories will offer rich resources for understanding social stereotype formation.

In summary, our work has demonstrated that not all items contribute to the gist equally in long-term memory, and sources of distinctiveness influence the formation of the gist. Our findings point to a disruptive role of attention and repeated exposure to gist memory formation and a potentially protective role from reward motivation. These findings provide insights into how humans extract summary statistics from individual experiences in long-term memory.

Methods

Experiment 1

Participants. In Experiment 1, we recruited 66 members of the University of Pennsylvania community (18–30 years old; normal or corrected to normal vision) to participate in the experiment for monetary compensation and course credit. The sample size was determined based on prior literature that reported an increased weight of particular items (Zeng et al., 2021). We excluded 8 participants because of low performance in gist memory (i.e. reported center error was larger than chance performance, defined by the distance between the center of the screen center and the center of the encoded items) and because of individual, gist performance, and the weight of the emphasized item out of 3 *SD* below or above average. Our reported results thus include 58 participants. All procedures were approved by the University of Pennsylvania IRB.

Stimuli and Procedure. Figure 3.1b illustrates the spatial locations of Experiment 1. The experimental procedure is displayed in Figure 3.1a. Each category (landmark of animal) was associated with a condition (reward or neutral). During the experiment, all participants completed two phases, training and testing. Each phase contained two cycles, a cycle for reward and another cycle for the neutral condition, with the order of the conditions randomized across participants. During the training phase, each cycle consisted of three rounds of training, where participants were trained to retrieve six locations of a category consecutively on a laptop. Each round of training consisted of two tasks, exposure and recall with feedback. In the exposure task, the locations appeared on

the screen one at a time, and participants clicked on each location in order to proceed. Consistent with prior studies using a similar paradigm (Zeng et al, 2021), on each trial, only one location was presented (never the full map), and the center of the encoded locations was never presented to participants. The instructions and the name of the locations were always displayed under the box of the locations to prevent possible influence on the memory of the locations. Figure 3.6 is a visual illustration of the exposure task.

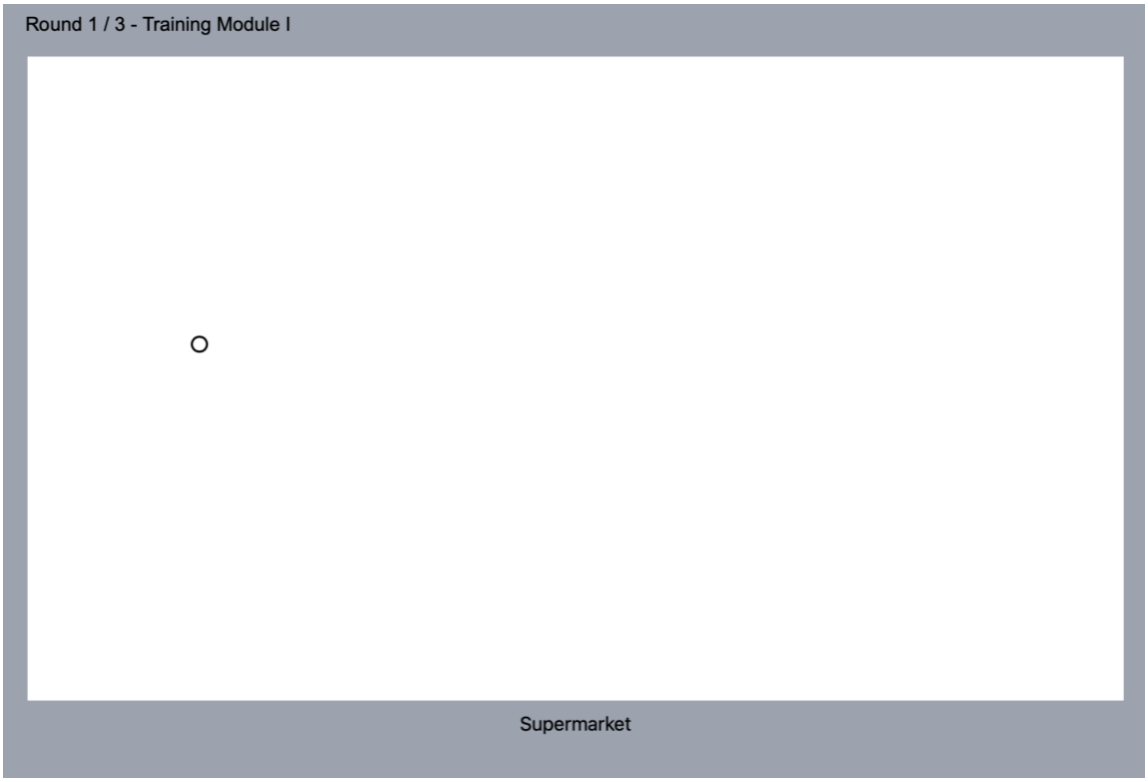


Fig. 3.6. Snapshot of the exposure task.

In the recall-with-feedback task, we asked participants to recall each location of a category by clicking on the screen when given its name as a cue, and feedback about their recall would be displayed: participants had one attempt to retrieve each location. If the

distance between the encoded location and recalled location satisfied the learning criterion (i.e. 60 pixels, which was defined by the minimum of the Euclidean distances between the pairs of the spatial locations), a message that their recall was correct would be prompted and the correct location would be displayed on the screen; otherwise, a message that their recall was incorrect would be prompted and the correct location would still be displayed on the screen.

The training of reward and the neutral cycle were identical except that in the reward cycle, the instruction explicitly said that remembering one item in the recall with feedback phase would work towards earning a bonus of $5 (which would double the payment for the experiment). Importantly, the appearance of this item was associated with signs of "$". The emphasized item was randomly chosen among the six items for each participant.

After participants completed the training phase of two cycles, they would be explicitly instructed that all the parts with monetary bonus were finished (this was intended to separate the influence of reward during encoding and recalling), but they were still encouraged to do their best in the next tasks. Participants then completed 10 unrelated arithmetic problems, which were designed to minimize potential influences from working memory. Finally, participants were tested on their gist and item memory, or in the reverse order. For the test of gist memory, they indicated their guess about the center of each category (e.g., 'Indicate the center (average location) of the landmarks you learned'). For the test of item memory, participants separately retrieved each location of the categories. The order of items was randomized in the tasks described above. All trials were self-paced. The total time for the experiment was approximately 10 min.

Error Measurement. We measured the accuracy for item memory (memory for each spatial location) and gist memory (reported center for each category of locations) as follows.

Item Memory Error (Dashed lines in Figure 3.2a): The error for each item was defined as the Euclidean distance between each retrieved location and its encoded location, where *d* (a, b) = sqrt((a1-b1)^2+(a2-b2)^2). Each participant's item memory error in each condition was computed as the average error for the six locations within each category. Chance performance based on the center of the encoded items was 133 pixels, which was determined by the average Euclidean distance between the center of encoded item locations and each encoded item location within each category. We average across the conditions. This distance corresponded to what participants' performance would be if they only remembered the center of item locations within each category and just clicked that category center when asked to recall any item.

Gist Memory Error (Solid lines in Figure 3.2a): The error for gist memory of each category/condition was defined as the reported center error, which was the Euclidean distance between the participant's reported center and the true center of all the encoded items for each category/condition. Chance performance was 249 pixels, which was the Euclidean distance between the center of the laptop screen and the true center of all encoded locations for each category/condition. We averaged across two conditions to get 249 pixels. This distance corresponded to what participants' performance would be if

they just clicked the center of the screen when asked to report the center for each category.

Weight Measurement. We computed the weight of the emphasized item on the gist memory as follows (Figure 3.7): For each condition, we projected the center participants reported on the connected line between the emphasized item participants reported and the average (center) of the reported neutral items with no reinforcement within the same category. The weight was then computed as the vector between the projected reported center and the average neutral items divided by the vector between the reported emphasized item and the average neutral items. Therefore, a higher weight value of an item indicates a higher contribution to the gist memory from that item. The value can range from negative to positive. When the value was ⅙, the item would contribute to the gist equally as other items, as the number of items within a category is 6.

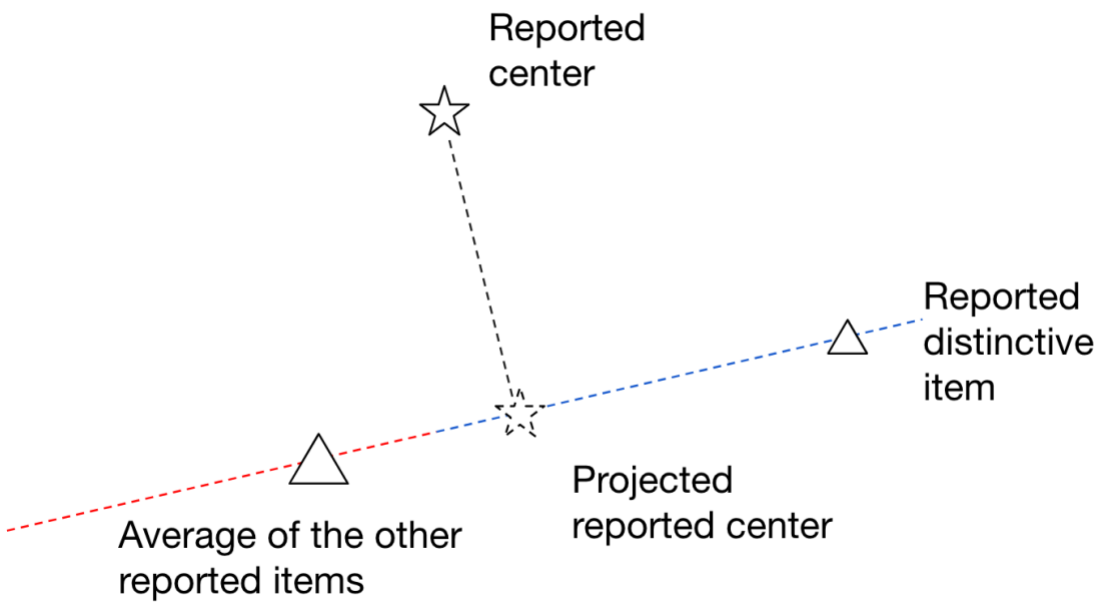


Fig. 3.7. Illustration of the weight computation. Red indicates when the distinctive item weighs less than other items and blue indicates when the distinctive item weights more than other items.

Statistics. To understand whether the item and gist memory varied between the neutral and the reward condition, we conducted paired t-tests between the two conditions for both the gist and item memory individually. To establish that participants were sensitive to the reward reinforcement, we used a linear mixed-effects model on item memory error with fixed effects of the type of the items (reward items in the reward conditions, neutral items in the reward conditions, and neutral items in the neutral conditions) with a random effect of the participant. In order to test whether there was a relation between item and gist memory, we used a linear mixed-effects model on gist memory error with fixed effects of item memory, the condition (reward vs. neutral), and their interaction, as well as a random effect of the participant ID. Finally, to examine whether the item with reward contributed more, the same, or less to the gist memory compared to other items, we compared their weight values against the weight assuming all items to be equal (i.e. ⅙) with Wilcoxon signed-rank tests, because the weight values were not normally distributed as determined by a Shapiro-Wilk test.

Experiment 2

Participants and stimuli. In Experiment 2, we recruited 712 participants through Amazon Mechanical Turk (MTurk) (located in the U.S.; HIT Approval Rate > 97%) for

monetary compensation. We conducted the experiment on MTurk to accelerate the data collection. The sample size was calculated based on the weight of Experiment 1, assuming $\alpha = .05$ and $\beta = 0.2$, which resulted in 84 participants per condition.

We thus generated 100 sets of stimuli for the yolked design with matching stimuli across groups. After participants signed up, we messaged them with an assigned ID and condition which they were required to enter at the beginning of the HIT. Through this information each participant entered, the program of the experiment generated the corresponding stimuli and procedures for them. Each set of stimuli contained six locations associated with landmark names (Figure 3.4b).

We excluded participants if they had low performance in gist memory (i.e. reported center error was worse than all of the participants in Experiment 1 and also larger than the distance between the center of the screen center and the center of the encoded items) and because of individual, gist performance, and the weight of the emphasized item out of 3 *SD* below or above average. We continued to collect data until we have around 100 participants per condition. Our reported results thus include 422 participants, with 112 participants in the reward condition (age: $M = 37$, $SD = 10$, gender: 42% Females), 102 participants in the repeat condition (age: $M = 38$, $SD = 12$, gender: 43% females), 108 participants in the attention condition (age: $M = 36$, $SD = 12$, gender: 57% females), and 100 participants in the neutral condition (age: $M = 40$, $SD = 12$, gender: 46% females). When we had a more relaxed exclusion criterion which does not exclude participants with gist performance worse than that in Experiment 1, we had 600 participants and the pattern of the results was the same. All procedures were approved by the University of Pennsylvania IRB.

Procedure. The training procedure for Experiment 2 was identical to that of Experiment 1 except that participants only experienced one condition with six locations (Figure 3.4) and also that there were two additional conditions, the attention condition, and the frequency condition, as described in the results section. In the reward condition, participants were told that remembering one particular item would work toward earning a monetary bonus. During training, the name of this item would have signs of "$". In the attention condition, participants were told that one item would be in a different color. During training, the location of that item would be displayed in red (Figure 3.8). In the repeated exposure condition, participants were told that one item would be associated with more training. During training, that item would appear four more times compared to other items in each task. In the neutral condition, no items were emphasized, but because the stimuli matched across groups, there would be one corresponding foil item that we could compare the emphasized item with.

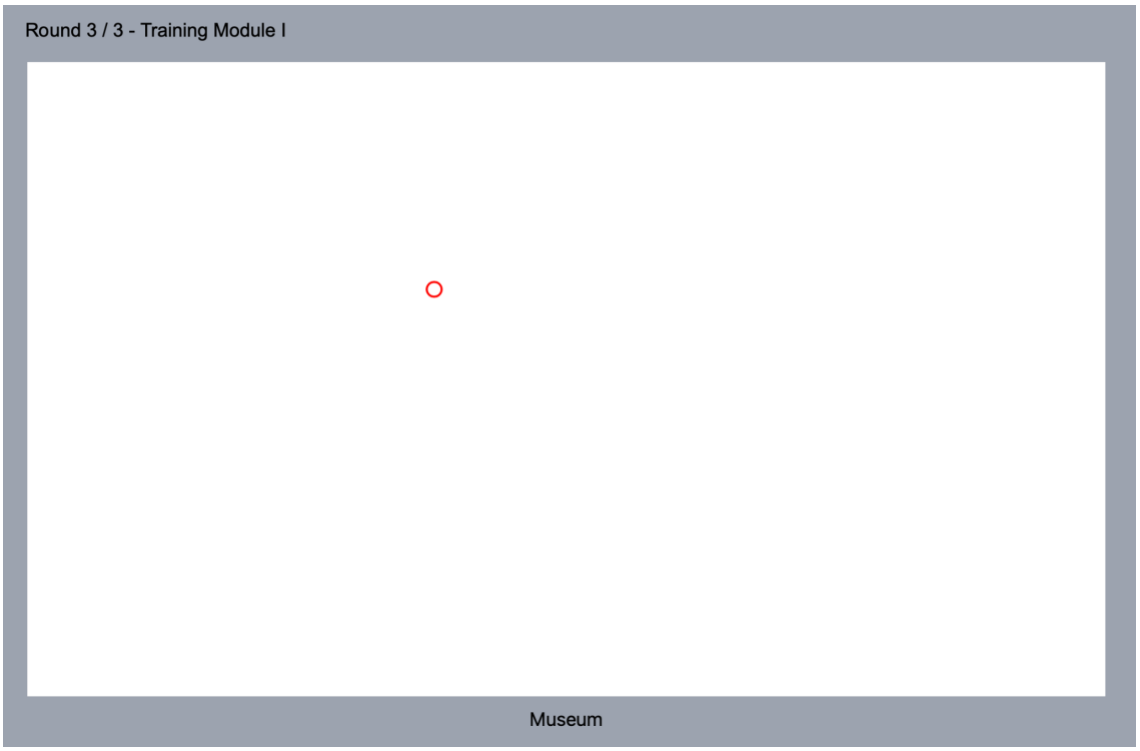


Fig. 3.8. Snapshot of the exposure task in the attention condition.

Error Measurement, Weight Measurement, and Statistics. The measurements for item memory error, gist memory error, and weight values are the same with Experiment 1. The statistics were thoroughly described in the main texts.

Supplementary Materials

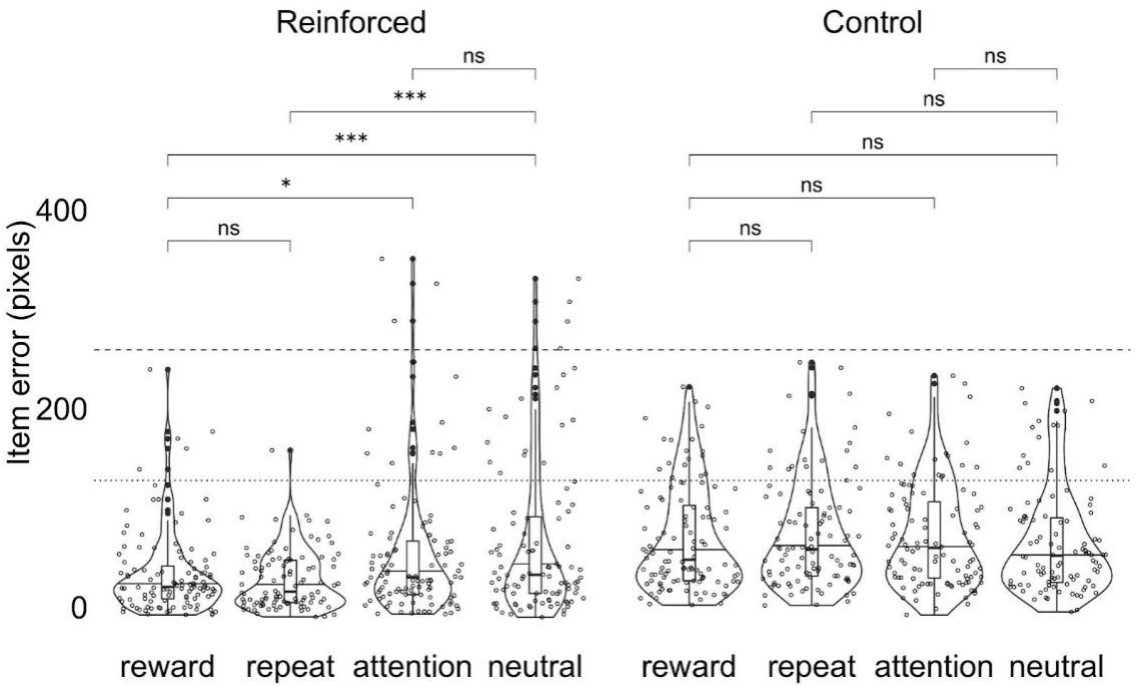


Fig 3.9. The error of participants' memory in the emphasized items and the neutral items by condition. Dashed lines indicate chance performance defined as the average distance between each encoded item location and the center of the screen. Dotted lines indicate the chance performance defined as the average distance between each encoded item location and the center of these encoded locations. For all figures, * indicates $p < 0.05$, *** indicates $p < 0.001$, and ns indicates $p > 0.05$ for the main effect of conditions between each pair of conditions by linear mixed-effects models controlling for stimuli. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outlier

Chapter 4: Negative memory bias in COVID-19 pandemic

Abstract

How might humans summarize their emotions of an experience that took place over an extended period of time? Would they differentially weigh the first day, the last day, the most typical day, or the most extreme day to evaluate the experience? It seems counterintuitive to think that someone would weigh an atypical day more. However, some research shows that the atypical moments (i.e. peaks) contribute uniquely to the estimation of the overall experience. In this study, 160 MTurk participants rated their daily emotions (e.g. happiness, stress, shock, anger) throughout the first two months of the COVID-19 pandemic. One week and one month later, participants recalled their average emotions over that two-months period as well as the date-specific emotions during these two months. We found a negative bias in both memories of the average emotions and memories of date-specific emotions. The recalled average emotions were more negative compared to the true average of the emotions. Moreover, the peak of negative emotions uniquely predicts this recalled average after controlling for other possible factors, such as the true average of these daily emotions. On the other hand, the recall of date-specific emotions was more negative than the actual date-specific emotions, but this negative bias decreased over time. We did not find the same effects for positive emotions. These findings provide new insights into the extraction of summary statistics in memory and emotion.

Introduction

Our memory of events is formed from moment-to-moment transient states of these events. For example, you probably remember how sad you were the day when your city locked down, and you probably remember how sad you were in general during the COVID-19 pandemic. How might humans summarize their emotions of an experience that took place over an extended period of time? Humans' ability to extract summary statistics from observing individual events has been widely examined in various fields of cognitive science, such as memory, perception, and concept formation. However, there is not enough evidence for the mechanism of summary statistics extraction in the memory of emotions.

Perception and memory literature shows that deviant instances disproportionally contribute to the summary statistics across these items. Long-term memory research shows that the extreme items may distort the memory of summary statistics more compared to other items (Richards et al., 2014; Zeng et al., 2021). It is unclear whether the cognitive principles in spatial memory apply to the domain of emotions, that is, whether the extreme emotions contribute more to the summary statistics of emotions in long-term memory.

Literature on the peak-end rule sheds light on this question by showing that the "peak" (i.e., the moment with the highest intensity) and the "end" of an experience, contribute to people's overall experience more than other factors such as the length of the experience (Ariely & Carmon, 2000). However, findings from this line of work are mostly from lab (Fredrickson & Kahneman, 1993). There is not much research investigating how

individual experiences of emotions contribute to the overall emotion in real life for an extended period of time. Moreover, the existing work that examines the effect of peak in real life with events spanning more than 1 week focuses on the evaluation of hedonic experiences, with the results being mixed. While some studies found evidence for peak predicting overall hedonic evaluation for a seven-day holiday at short delays but not long delays (Geng et al., 2013), other studies did not find the peak or trough (the lowest points) to be good predictors for overall hedonic evaluation for a 32-day winter holiday or events in high school (Kemp et al., 2008; Kemp et al., 2012).

In addition, little is known about whether there may be a difference between positive and negative emotions in the contribution of peak to the overall emotions in long-term memory. Prior literature on perception shows positive emotions facilitate global processing while negative emotions facilitate local processing in summary statistics extraction (Fredrickson & Branigan, 2005; Peng et al., 2022). However, there is not much evidence for this difference in long-term memory, possibly because of a lack of opportunity to measure and compare the two within the same context.

These findings point to a gap in the current literature about how extreme emotions contribute to the evaluation of the summary statistics of the emotions in long-term memory. Will the memory of overall emotions be distorted toward the peak, that is, the extreme emotions, over a long extended period of time? Does this peak effect vary by evaluation of positive and negative emotions? It will be important to examine this effect after controlling for other factors of the daily emotions, such as the true average, the first day, the last day, and participants' current emotions.

Moreover, the peak effect may lead to certain biases in memory. For example, if the worst moment during the COVID-19 pandemic contributes more to our memory of average sadness, then our recalled average sadness may be even more negative compared to the true average of our daily sadness. Alternatively, if the best moment during the pandemic contributes more to our memory of average sadness, then our recalled average sadness may be less negative compared to the true average of our daily sadness. Prior research on emotion memory biases provides mixed results. Some studies found positive biases and some found negative biases (Adler & Pansky, 2020; Baumeister et al., 2001). A possible explanation for the discrepancy is that the effect of emotions on memory varies over time. For the purpose of survival, humans need to remember negative events well in the short term to make sure they can deal with the negative events in time in order to survive. However, for the purpose of maintaining well-being and protecting their mental health, humans will need to prioritize positive information in the long term (Adler & Pansky, 2020). Moreover, the existing studies focus on the recall of individual events rather than the recall of summary statistics. It will be useful to seek evidence for the change of emotional memory bias over time for both the memory of summary statistics and the memory of individual events. In order to understand the possible change of positive and negative bias over time, we collected participants' recall of summary statistics and date-specific emotions at a delay of one week and one month after the daily surveys.

The unfortunate COVID-19 pandemic provides a rare opportunity for us to investigate how humans summarize their positive and negative emotions from a long and emotionally varied event in real life as well as the change of memory bias over time. We

selected the first two months of the pandemic as the targeted time window based on the belief that our participants' emotions about COVID-19 would be more varied in the early phase of the pandemic. Through daily surveys across two months to participants on MTurk, we collected 84 participants' daily general happiness and negative emotions intensity on COVID-19. Through final surveys at a delay of one week and one month after these daily surveys, we collected participants' recalled average emotions for general happiness and negative emotions on COVID-19. We explored what factors of the daily emotions (e.g., the first day, the last day, the peak, participants' current emotion) contribute to the recall of their average emotions. We examined whether participants' recall of their average emotions would be negatively or positively biased compared to the true average of their daily emotions and whether the bias changed over time. Our study sheds light on how humans summarize their daily emotions to form an overall evaluation of their emotions over an extended period of time.

Methods

Participants

In order to have a statistical power of t-tests that can detect a day-to-day emotional change of 20% with $\alpha = .05$ and $\beta = 0.2$ and correlation analyses of $\alpha = .05$, $\beta = 0.2$, and $r = 0.3$, the sample size was determined to be 85. We started collecting data on Amazon Mechanical Turk (www.mturk.com) for 160 participants who were in the United States with a HIT approval rate > 95, with the anticipation that half of the participants would continue to complete the follow-up surveys. Participants were recruited from the United

States. Finally, we have data for 84 participants (48% Male, 51% Female, and 1% Other) who have continued to complete the daily surveys and the final surveys (Mean: 44 days, *SD*: 17 days).

Procedure

Daily Surveys: On every day from March 16 to May 11, 2020, participants reported their emotions about COVID-19 on a Qualtrics questionnaire (Figure 4.1; blue, top). Specifically, the question for general happiness was "In general, how happy do you feel today?" on a scale of 1 to 7 (decimal numbers are allowed). The question for negative emotions was: "At this moment, how strong or intense are your feelings about the spread of coronavirus?" on a scale of 1 to 7 (decimal allowed), with 7 emotions to rate (sad, angry, fear, frustration, confusion, shock, and stress). These questions were embedded in other questions related to COVID-19 such as participants' knowledge of COVID-19. The time period was determined for the consideration of allowing variability of emotions across days.

Final Surveys: One week and one month after March 11, 2020, participants reported their memory of their average emotions about COVID-19 during the time they received the daily surveys (Figure 4.1; purple, middle). Specifically, the question for general happiness was "On average, how happy do you think you have been in general in the eight weeks between March 16 and May 11?" on a scale of 1 to 7 (decimal numbers are allowed). The question for negative emotions was: "On average, how strong or intense do you think your feelings have been about the spread of coronavirus in the eight weeks

between March 16 and May 11?" on a scale of 1 to 7 (decimal allowed) with 7 emotions to rate as in the daily surveys. These questions were embedded in other questions about COVID-19 to match the structure of the daily surveys. In order to retrieve participants' emotions when they reported the average so that we can control for it in our analyses, we asked participants to report their general happiness and negative emotions intensity in the final surveys as in the daily surveys.

In order to retrieve participants' memory of date-specific emotions, we asked participants to fill out questions about memories of their date-specific emotions (Figure 4.1; dark green, bottom). Participants recalled four personal events and four news headlines between March 16 and May 11, with the order of the two kinds of memories randomized. After the recalls, participants reported their memory of their general happiness and negative emotions on the day these events happened on a scale of 1 to 7 (decimal allowed). In order to retrieve the specific dates for these emotions, at the end of the survey, we asked participants to identify the specific date these personal events happened and the specific dates they learned about the news as accurately as possible. They were encouraged to look up from their record or search on the web to be as accurate as possible.

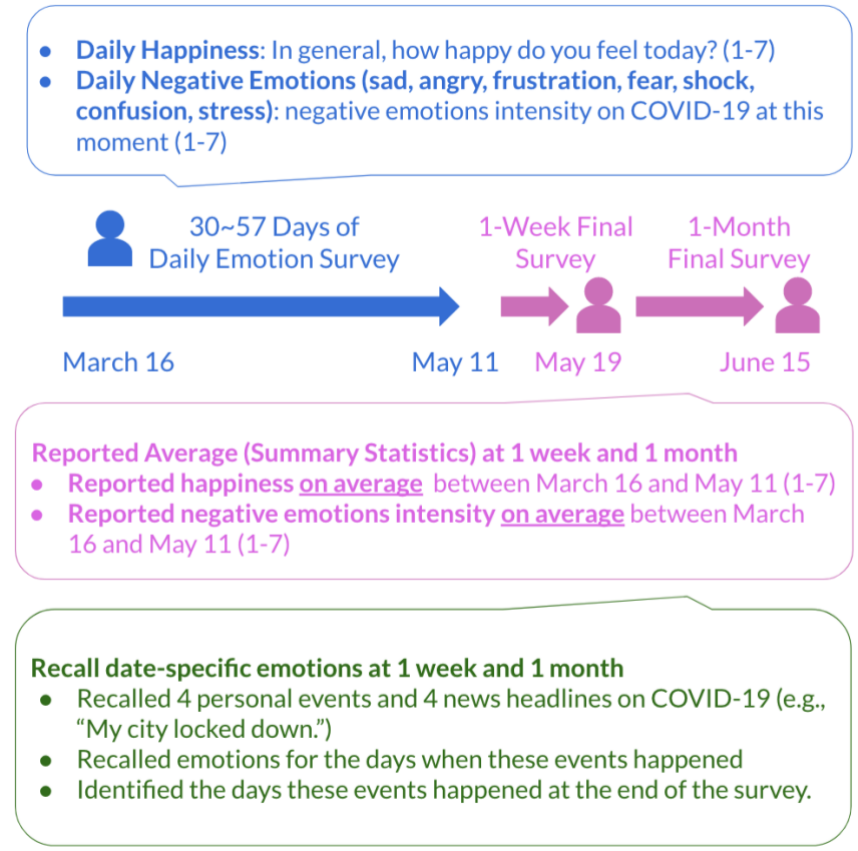


Fig. 4.1. Schematic illustration of the timeline and the procedure of the surveys. Blue (the top) indicates daily surveys. Purple indicates the questions about the overall emotions in the final surveys. Green (the bottom) indicates questions for date-specific emotions in final surveys.

Results

Because we did not have a prediction for the difference between the negative emotions, we combined the intensitive of negative emotions by averaging across all the negative emotions for the daily emotions, average emotions, and the emotions participants experienced now when reporting the average. Figure 4.2 visualizes a typical participant's composite emotions reported in their daily surveys, the "reported

average”(i.e., the average emotions participants reported in their final surveys), and the emotions participants had when filling out the final surveys.

As visualized in Figure 4.2, we can thus identify participants’ emotions on the first day in the two months, their emotions on the last day in the two months, the “peak” of their emotions (i.e., the composite negative emotions and the general happiness at the highest level), and the “trough” of their emotions (i.e., the composite negative emotions and the general happiness at the lowest level) from the emotions reported in the daily surveys. Importantly, we computed the “true average” (i.e., the average of the daily emotions of a participant) from the daily emotions.

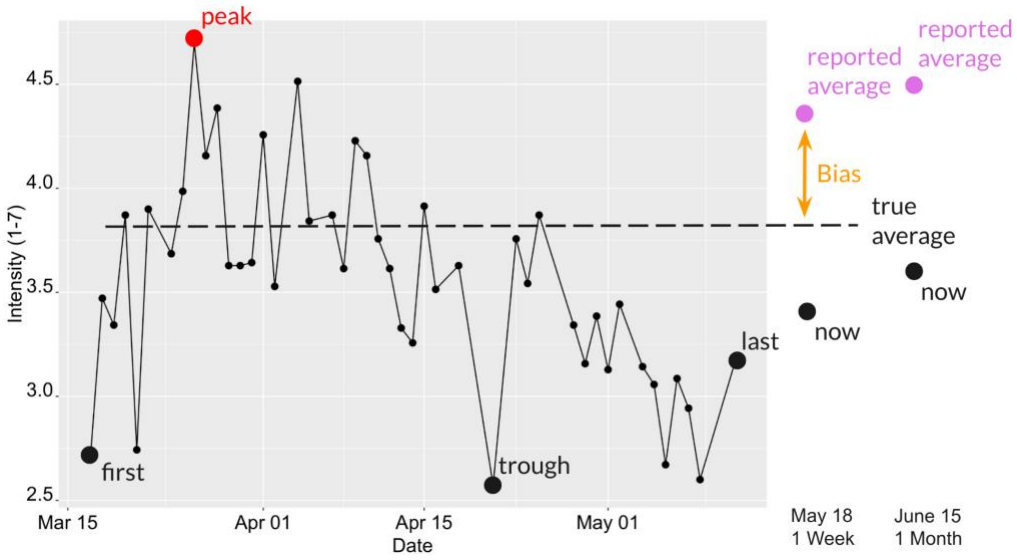


Fig. 4.2. Composite average of 7 negative emotions for a typical participant

Participants’ daily emotions significantly changed over time in the first two months of the COVID-19 Pandemic

In order to examine whether there is a day-to-day change in participants' emotions for the main analysis, we used a linear model to evaluate the effects of dates on daily general happiness and negative emotions. We found that the date significantly predicted the composite negative emotions, $SSE = 99.3$, $F(1, 3620.2) = 277.59$, $p < 0.001$, and significantly predicted the general happiness, $SSE = 26.3$, $F(1, 4781.5) = 49.93$. These results indicate that participants' positive and negative emotions significantly changed across days in the first two months of the COVID-19 pandemic.

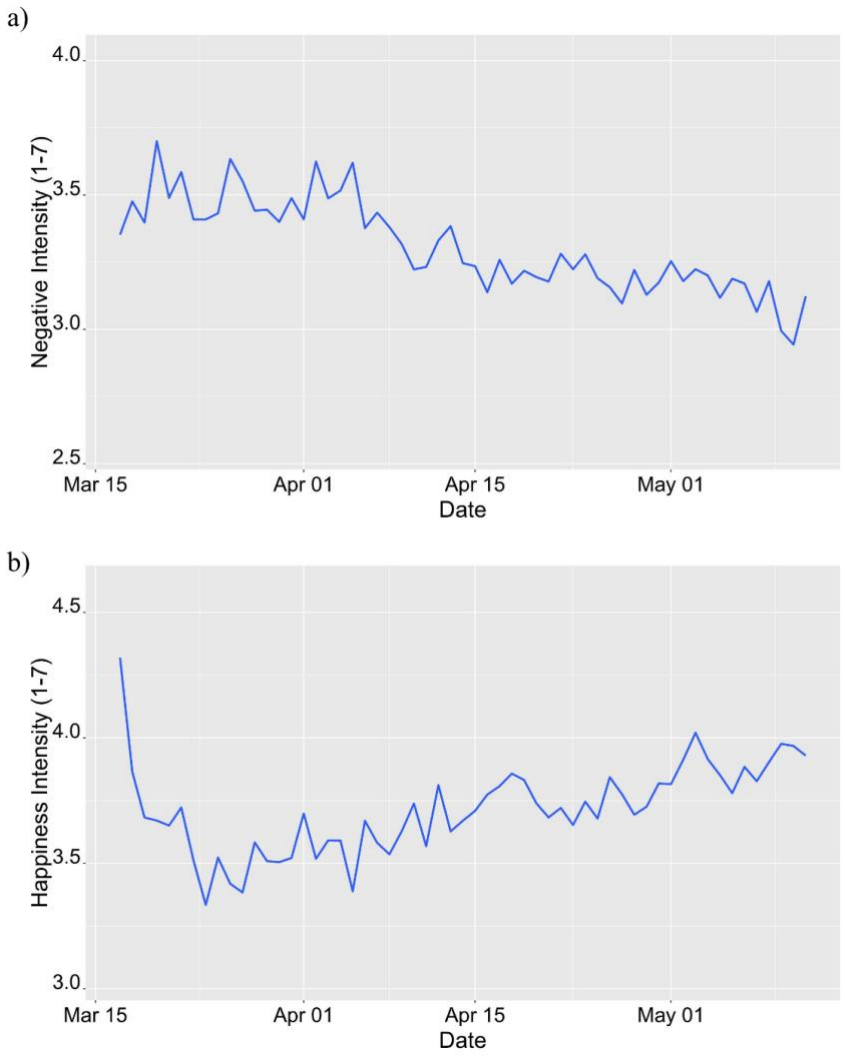


Fig. 4.3. Participants' emotion intensity for negative and positive emotions both changed overtime. a) Intensity of negative emotions composite average across participants over time. b) Intensity of general happiness average across participants over time. Higher values indicate higher intensity.

Recalled average emotions accurately reflect true average emotions, but they are negatively biased

In order to determine whether participants' reported average emotions reflect their true average emotions, we averaged the 1-week and 1-month reported average because they highly correlated with each other and computed the correlation between this reported average and the true average of participants' daily emotions. The analyses showed a significant positive correlation between the true average and the reported average for their negative composite emotions, $r(82) = 0.85$, $p < .001$ (Figure 4.4), and their general happiness, $r(82) = 0.87$, $p < .001$, which suggests that participants' reported average emotions accurately reflect their true average emotions.

Moreover, the bias (recalled average of participants' composite negative emotions - the true average of these emotions) was significantly higher than 0 (Wilcoxon signed-rank tests: $Z = 2.69$, $p = 0.007$). On the other hand, we did not find the same bias for general happiness (Wilcoxon signed-rank tests: $Z = 0.32$, $p = 0.749$). Taken together, these results suggest that although participants' memory of their average emotions accurately reflects the true average of their daily emotions, this memory was negatively biased.

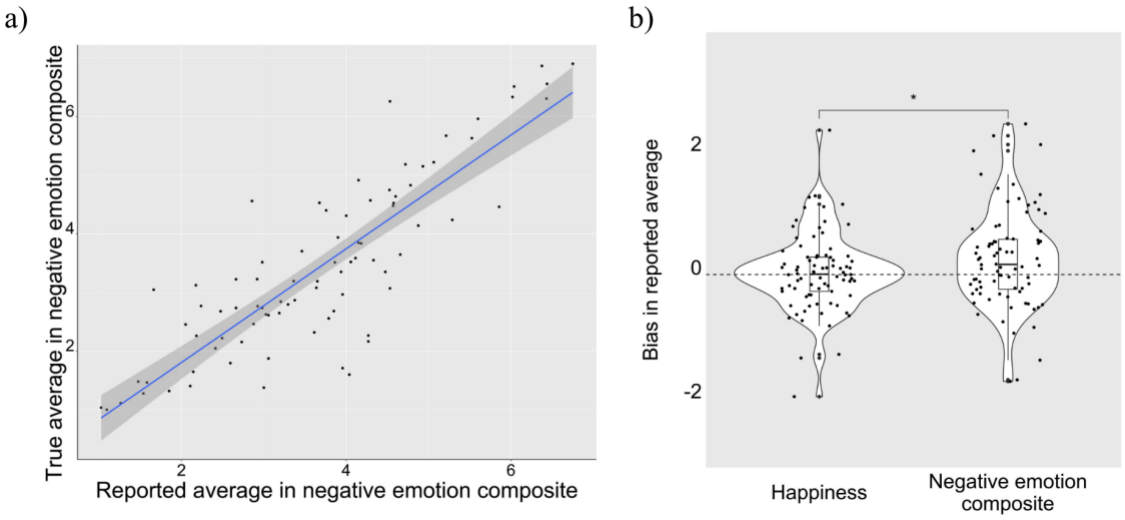


Fig. 4.4. Participants' recall of average emotions accurately reflects the true average of their daily emotions, but this recall is negatively biased. (a) Participants' reported average in the composite negative emotions is significantly correlated with the true average in their composite negative emotions. (b) Participants' recalled average for their composite negative emotions is more intense compared to the true average of these negative emotions. For happiness, participants' reported average is not different from the true average. * indicates $p < 0.05$ by Wilcox test. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers.

The peak significantly predicts the reported average even after controlling for other predictors for negative emotions, but not for happiness.

In order to understand what factors contribute to participants' memory of average emotions, we fit a linear model on participants' memory of average emotions with fixed effects of true average across daily emotions ("true average"), their emotions at the time

they reported the average ("now"), the peak, the trough, the first day, and the last day of the daily emotions (Figure 4.2). For both general happiness and negative emotion composite, we first added true average and then added now as fixed effects, after which we added peak as a fixed effect (Table 1). For negative emotions, we found a main effect of peak, $SSE = 6.15$, $F(1, 80) = 22.15$, $p < 0.001$ after controlling for true average and now (Model 3 in Table 1). The effect of peak persisted after adding other covariates of first, last, and trough to the model (Model 4 for Composite Negative Emotion in Table 1). These results suggest that for negative emotions, the peak significantly predicts memory of average emotions after controlling for other days of emotions.

However, we did not find the same effect on happiness. For happiness, we did not find the main effect of peak, $SSE = 0.05$, $F(1, 80) = 0.17$, $p = 0.682$ after controlling for true average and now (Model 3 for general happiness in Table 1). Similar to negative emotions, we then added covariates of first, last, and trough to the regressions model as fixed effects (Model 4 for General Happiness in Table 1). The model showed an effect of the first day, $SSE = 1.64$, $F(1, 60) = 7.18$, $p = 0.009$.

Table 1. Regressions for the recall of average emotions

	Composite Negative Emotion				General Happiness			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
True Average	0.751 [***]	0.330 [***]	-0.065 [ns]	0.071 [ns]	0.867 [***]	0.501 [***]	0.510 [***]	0.383 [*]
Now		0.540 [***]	0.620 [***]	0.663 [***]		0.450 [***]	0.461 [***]	0.360 [***]
Peak			0.365 [***]	0.411 [***]			-0.031 [ns]	-0.090 [ns]
Trough				0.040 [ns]				0.012 [ns]
First				-0.058 [ns]				0.189 [**]
Last				-0.174 [ns]				0.376 [ns]
Adjusted *R*-Square	0.724	0.801	0.841	0.858	0.761	0.822	0.821	0.855
F	219.2	167.5	148.2	68.2	264.6	193.1	127.5	59.1
p-value	***	***	***	***	***	***	***	***

*** Denotes significance at $p < 0.001$

To further contrast general happiness and negative emotions, we fit a mixed-effects linear model on participants' memory of average emotions with all fixed effects (true average, now, peak, trough, first, and last) and an interaction term of the type of emotion (happiness vs. negative emotions), and a random effect of participants' ID. The main effect of emotion type is not significant, $SSE = 0.001$, $F(1, 121) = 0.005$, $p = 0.94$. The only significant main effects are now, $SSE = 9.66$, $F(1, 121) = 38.93$, $p < 0.001$, and peak, $SSE = 1.50$, $F(1, 121) = 5.98$, $p = 0.015$. Importantly, there is an interaction between emotion and peak, $SSE = 3.62$, $F(1, 121) = 14.60$, $p < .001$, and an interaction between emotion and first day, $SSE = 1.36$, $F(1, 121) = 5.49$, $p = .02$. These results suggest that the factors that contribute to participants' memory of average emotions are similar but different for general

happiness and negative emotions. In particular, how participants feel now contributes to participants' recall of average emotions for both positive and negative emotions. The peak contributes more to participants' memory of the average negative emotions compared to general happiness. The first day contributes more to participants' memory of average general happiness compared to negative emotions.

Negative bias in recalled date-specific emotions decreased over time

In order to understand how summary statistics may influence participants' recall of date-specific emotions, that is, whether participants' recall will bias towards the summary statistics over time (Zeng et al., 2021), we calculated the error of participants' date-specific emotions ("bias") by subtracting their recall for the emotions at the day and their emotions reported at that day.

Overall, the bias in participants' recalled date-specific negative emotions was significantly higher than 0, Wilcoxon signed-rank tests: $Z = 3.03$, $p = 0.002$ (Figure 4.5a), which suggests participants' recall of date-specific emotions was negatively biased. There was not a significant difference between the recall bias at 1 week or at 1 month for both types of events (ps > 0.74). In order to further understand the change of negative bias over time, we fit a mixed-effects linear model on participants' recall bias with fixed effects of delay in days and an interaction term of the memory type (personal events vs. news), and a random effect of participants' ID. The date of the recall, which is the delay, significantly negatively predicts the recall error, $SSE = 4.87$, $F(1, 306.12) = 5.37$, $p = 0.021$ (Figure 4.5b). Other effects were not significant (ps > 0.47). These results suggest that although participants' recalled date-specific emotions were negatively biased at both time points,

this negative bias decreased over time.

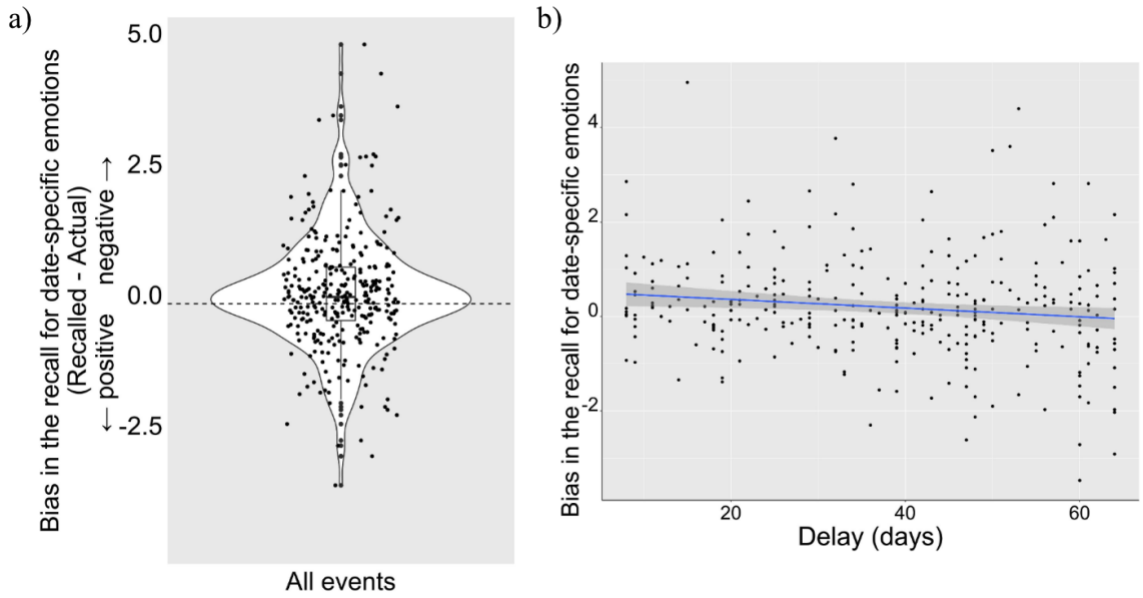


Fig. 4.5. Negative bias in the recall of date-specific emotions. (a) Bias in the date-specific emotions participants reported for both personal and news events. The band indicates the median, the box indicates the first and third quartiles, the whiskers indicate ± 1.5 × interquartile range, and the solid points indicate outliers. (b) The negative relation between days of delay and the bias for date-specific emotions.

On the other hand, although participants' recalled date-specific general happiness was also significantly lower than 0, Wilcoxon signed-rank tests: $Z = 6.50$, $p < 0.001$, we did not find the same relationship between the delay and the recall error, $SSE = 1.49$, $F(1, 265.76) = 0.79$, $p = 0.37$. Although we found a decrease in emotional memory bias over time for negative emotions, we did not find evidence for the same change in negative bias for general happiness.

Discussion

We examined how humans extract the summary statistics of emotions from individual experiences through two months of surveys at the beginning of the COVID-19 Pandemic. We demonstrated a negative bias in the summary statistics extraction of emotions. In particular, for the daily negative emotions, the summary statistics were recalled to be more intense compared to the true average. Critically, the peak of the daily emotions uniquely contributes to the summary statistics for the negative emotions. On the other hand, memory for the individual events, which are the recalled date-specific emotions, was also recalled to be more negative compared to the true date-specific emotions, but this bias decreased over time. We think that the cognitive mechanism of summary statistics functions differently for positive and negative emotions.

Our work integrates research in peak-end and research in long-term memory research (Richards et al., 2014; Zeng et al., 2021) by designing a paradigm that bridges the two lines of research and tests the role of extreme items in the formation of summary statistics. Our results that the peak uniquely contributes to the average negative emotions are consistent with prior literature on the peak-end rule (Ariely & Carmon, 2000). Although the surveys in this study do not cover the full scope of COVID-19 and therefore do not really contain an end, this study adds to this line of prior work by demonstrating the unique role of the peak in evaluating the summary statistics of emotions over an extended period of two months, whereas the few prior studies of the longest period only cover a week to a month and only tests hedonic experiences (Geng et al., 2013; Kemp et al., 2008). Due to the pandemic, our study was able to demonstrate that the role of the

peak may be different for positive and negative experiences in the same comparable context and examine the role of the peak in emotions in an influential event in the world. On the other hand, unlike past work in the peak-end rule, our study asked participants to report "average" emotions instead of "overall" emotions, which transforms the paradigms in the peak-end rule to fit the paradigms in long-term memory (Zeng et al., 2021). Our current results benefit the long-term memory research literature by demonstrating that the cognitive principle that extreme items contribute more to the summary statistics applies to the domain of emotions, and maybe negative emotions only.

A few limitations of the current study should be noted about the contrast between general happiness and negative emotions. First, the questions about general happiness were embedded in a questionnaire on COVID-19. This context may encourage a negative bias in recalling general happiness. A perfect control condition would be having another group of participants fill out a series of surveys with the general happiness questions alone during the early pandemic. Although it was impossible to return to the early pandemic and conduct this control condition, the current results are consistent with findings from some prior research on happiness Researchers tracked participants' happiness during a week of vacation, measured the overall happiness during this time, and did not find peak-end effects on the overall happiness (Geng et al., 2013; Kemp et al., 2008). Moreover, by having the current survey design, we allowed for a matched context between general happiness and negative emotions. If it was the context of COVID-19 questions that determined participants' responses for all emotions, we should expect to see the same effect for general happiness and negative emotions. For example, the "trough" of happiness should uniquely contribute to the average happiness. However, we did not observe this effect. Another

limitation is that the wording of the survey questions about general happiness do not match the questions about the negative emotions, which was due to a consideration of avoiding unnatural questions that ask participants how happy they are about coronavirus. It was unclear how the different phrasing will influence the results. However, if we assume the cognitive mechanisms apply to the extraction of summary statistics of all emotions to the same extent. Our results, at the very least, show that the cognitive mechanisms of summary statistics extraction do not apply to all emotions equally.

We found differences between the extraction of summary statistics for positive and negative emotions and these results add to the literature on emotions. The peak influences the overall emotions of negative emotions more than positive emotions, which provides evidence for the theory that positive emotions facilitate global processing and negative emotions facilitate local processing in long-term memory (Peng et al., 2022). Moreover, our finding that emotions participants feel now contribute to the recalled average for both general happiness and negative emotions adds new evidence to the memory reconstruction framework in emotions, which proposes that memory of emotions is based on an inherently reconstructive process and should be susceptible to biases for adaptive values (Adler & Pansky, 2020). In particular, participants' current emotions play a special role in the reconstructive process of their past overall emotions. In addition, our analysis showed that the negative bias for the memory of date-specific emotion decreased over time, which added new evidence to the adaptive framework that positive memories should be prioritized over time to maintain well-being (Adler & Pansky, 2020).

In summary, we have shown a negative bias in both memories of overall emotions and memories of date-specific emotions that lasted over time, and that the negative

emotion of highest intensity may uniquely contribute to the overall recall. However, the negative bias for recalling date-specific emotions diminished over time. Moreover, we found that positive and negative emotions vary in the extraction of summary statistics as well as the change of bias in time. These findings provide new insights into the extraction of summary statistics in memory and emotion.

General Discussion

Our research explored the cognitive mechanisms of summary statistics across fields in psychology, from memory and consolidation to emotion and motivation. Chapter 2 used a spatial memory paradigm to demonstrate that the memory of summary statistics, initially extracted from items, starts to bias memory for the items over time. A deviant item changes the summary statistics extraction. Chapter 3 used the same paradigm to further disentangle how different sources of item distinctiveness influence summary statistics extraction. Chapter 4 extends this examination of summary statistics, especially the influence of deviant items, to an impactful event in real life that lasts over an extended period of time, COVID-19, and demonstrates a bias on summary statistics that may vary by type of emotions.

There are generalities and specificities across various fields in psychology. For example, Chapters 2 and 4 both show that extreme items uniquely contribute to the summary statistics of the items in the lab and real-life events. Chapter 2 shows that an outlier location that is deviant from others locations shifts the summary statistics of these locations. Chapter 4 extends this rule to summary statistics of emotions during COVID-19 by demonstrating a unique contribution from extreme emotions to the summary statistics of the emotions for negative emotions. When evaluating the summary statistics over time, the extreme items uniquely contribute to the summary statistics both in memory of spatial locations and memory of emotions.

On the other hand, our results also suggest specificities in these fields and potential interactions between domains in cognition. Although in Chapter 4 we found that the peak

of emotions uniquely contributes to the summary statistics for negative emotions, we did not find the same peak effect for positive emotions. Moreover, Chapter 4 detected a negative bias in the summary statistics for negative emotions, but not positive emotions. These results suggest that emotions may play a role in the extraction of summary statistics, which potentially link to the adaptive values of memory to maintain well-being. Chapter 3 manipulates the distinctiveness of the items and discovered that reinforcing an item with various sources of distinctiveness will change its contribution to the summary statistics. Despite similar levels of improved accuracy of the emphasized item, repeated exposure distorted the gist memory by increasing the weight of this emphasized item, whereas the reward did not. It is possible that the reward, through facilitating the same neural network as summary statistics extraction, improves the integration across items, balancing their weights, and thus protects the gist from being distorted. Alternatively, the increased weight with repeated exposure may be due to the interferences from repeated exposure on the processing of the summary statistics. This increased weight may be specific to repeated repetition. Work can be done to continue to explore the factors that will influence the weights of items to deepen our understanding of the mechanism of summary statistics extraction.

In addition, working memory and long-term memory traditionally have been studying summary statistics extraction separately, but their findings reveal similar phenomena, such as memory of summary statistics being retained while memory for items being discarded. Our work connects these two fields by creating experimental paradigms that share commonalities in both fields. For example, Chapters 2 and 3 used a spatial memory paradigm that allowed for weight computation for particular items, similar to ensemble

perception literature (Whitney & Yamanashi, 2018). Chapter 2 found an increased gist-based bias on the memory of items over time. This bias in long-term memory mirrors the gist-based bias in working memory research and suggests that the idea of a hierarchical organization of memory may apply in long-term memory consolidation: as the uncertainty of item memory increases, they are more biased towards the gist to improve the accuracy of item memory retrieval (Brady and Alvarez, 2011; Lew and Vul, 2015; Orhan and Jacobs, 2013).

Moreover, in working memory literature, outlier items tend to be discarded in summary statistics (de Gardelle and Summerfield, 2011; Haberman & Whitney, 2010). Chapter 2 discovered that the outliers contribute more to the explicit recall of gist memory, but the outliers were discarded in the implicit center that biased item memories over time. This implicit center that discarded outliers may reflect a sampling strategy in long-term memory consolidation that is similar to the working memory literature (de Gardelle and Summerfield, 2011; Haberman and Whitney, 2010). On the contrary, the gist that overweighted the outlier participants reported may be associated with a sampling strategy that is explicit and different from the gist extraction in working memory literature.

To be more specific, when there are outliers in the group, the human mind may automatically disregard the outlier and compute the summary statistics of the "median" which excludes the outlier. This disregard is fast, effortless, and free from the need for a deliberate recall of the individuals. This may account for what is found in the ensemble literature and the gist-based bias center we found in long-term memory. On the contrary, when there is an outlier but participants are asked to compute the average, which is the

case in Chapter 2, participants have to consciously recall the outlier and incorporate this outlier with the median to get the average, which shifts the weight of the outlier in the average computed. This may account for the explicitly recalled center that overweighs the outlier in Chapter 2. Our results offer a start for future research endeavors to further understand the summary statistics extraction across ensemble perception and long-term memory.

Why would the mind automatically compute the mean as opposed to the average when there is an outlier? This mechanism may be of adaptive value. Imagine you are looking for food in the forest. You have the memory that the food usually can be found at certain locations, and you also have the memory that sometimes food shows up at outlier places that are far from the majority of locations. Today, you need to look for food again. Doesn't it make more sense to disregard the outlier locations and check first at the majority of locations where food usually is? This is similar to the median computation discarding the outlier. Only when you cannot find the food at the majority of locations will it make sense for you to resort to the outlier location. This conscious recall of the outlier will then shift the weight of the outlier.

Will different decision processes involve different summary statistics computations (mean vs. median)? When will it make more sense to compute the mean, as opposed to the median? Sometimes the mean is fairer. For social categories, the default median computation can lead to a bias. If people always disregard the outlier when forming an overall gist of a social category, when most of the individuals in this social category are associated with unfavored characteristics, the occurrence of an outlier with favored characteristics will not change people's impression of the group, even though a mean

with the outlier taken into account is a more accurate reflection of the group summary statistics. Understanding this mechanism of summary statistics extraction may help us fight this bias in society.

Our current findings provide other interesting implications for stereotype formation in the social category. For example, Chapters 2 and 3 both discovered the distinctiveness of individual items will influence their contribution to the summary statistics. Chapter 2 provides evidence that the "gist" starts to stabilize and bias memory for individuals after an extended period over time. Work in social stereotype formation suggests that after a stereotype has been formed from accumulating experiences from individuals, judgment on the group may be made from the abstract information independently of the individuals (Sherman et al., 2013). It is still unclear how the experiences of the individuals will shape the stereotype and the bias on social categories, and how these memories evolve over time in society as it was hard to acquire a "blank" social group that participants never had any experience with. Taking the question of gist memory formation to a greater scale, it remains to be discovered how gist memory passes across generations over time to become culture and history and shapes our personalities and behaviors (Schulz et al., 2019). Our lab work offered a starting point to integrate the cognitive mechanism of summary statistics extraction into a broader application of understanding gist memory formation in society.

Taken together, these three chapters enlighten our understanding of how the human mind integrates across individual experiences to form summary statistics over time by bridging fields ranging from perception and memory to emotion and motivation. Our work suggests possible generalities and specificities between these fields. They open up

new possibilities for integrating theories for summary statistics extraction across the fields, offer a promising start to deepen the research endeavor in summary statistics extraction in human cognition, and provide implications and insights into understanding human memory in society.

Reference

Adler, O., & Pansky, A. (2020). A “rosy view” of the past: Positive memory biases. *Cognitive Biases in Health and Psychiatric Disorders*, 139–171. https://doi.org/10.1016/c2018-0-00401-6

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Apicella, C. L., Little, A. C., & Marlowe, F. W. (2007). Facial averageness and attractiveness in an isolated population of hunter-gatherers. *Perception*, *36*(12), 1813–1820. https://doi.org/10.1068/p5601

Ariely, D., & Carmon, Z. (2000). Gestalt characteristics of experiences: The defining features of summarized events. *Journal of Behavioral Decision Making*, *13*(2), 191–201. https://doi.org/10.1002/(SICI)1099-0771(200004/06)13:2<191::AID-BDM330>3.0.CO;2-A

Ariely, D., & Carmon, Z. (2003). *Summary assessment of experiences: the whole Is different from the sum of its parts*. Retrieved from https://www.ebsco.com/terms-of-use

Armstrong, S., Gleitman, L., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*, 263–308.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad Is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Chun, M. M., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology*, *17*(2), 177–184. https://doi.org/10.1016/j.conb.2007.03.005

Clewett, D., & Murty, V. P. (2019). Echoes of emotions past: How neuromodulators determine what we recollect. *ENeuro*, *6*(2). https://doi.org/10.1523/ENEURO.0108-18.2019

Corbin, J. C., & Elizabeth Crawford, L. (2018). Biased by the group: Memory for an emotional expression biases towards the ensemble. *Collabra: Psychology*, *4*(1), 1–8. https://doi.org/10.1525/collabra.186

Cowan, E. T., Schapiro, A. C., Dunsmoor, J. E., & Murty, V. P. (2021). Memory consolidation as an adaptive process. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-01978-x

De Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United*

States of America, *108*(32), 13341–13346. https://doi.org/10.1073/pnas.1104517108

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition and Emotion*, *19*(3), 313–332. https://doi.org/10.1080/02699930441000238

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*(1), 45–55. https://doi.org/10.1037/0022-3514.65.1.45

Fredrickson, B. L., & Levenson, R. W. (1998). Positive Emotions Speed Recovery from the Cardiovascular Sequelae of Negative Emotions. In *Cognition and Emotion* (Vol. 12). https://doi.org/10.1080/026999398379718

Fu, Y., Zhou, Y., Zhou, J., Shen, M., & Chen, H. (2021). More attention with less working memory: The active inhibition of attended but outdated information. *Science Advances*, *7*(47), 1–14. https://doi.org/10.1126/sciadv.abj4985

Geng, X., Chen, Z., Lam, W., & Zheng, Q. (2013). Hedonic evaluation over short and long retention intervals: the mechanism of the peak-end rule. *Journal of Behavioral Decision Making*, *26*(3), 225–236. https://doi.org/10.1002/bdm.1755

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*(1), 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Goldstone, R. (1994). Influences of Categorization on Perceptual Discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178

Graves, K. N., Antony, J. W., & Turk-Browne, N. B. (2020). Finding the pattern: On-line extraction of spatial structure during virtual navigation. *Psychological Science*, *31*(9), 1183–1190. https://doi.org/10.1177/0956797620948828

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception & Psychophysics*, *72*(7), 1825–1838. https://doi.org/10.3758/APP.72.7.1825.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202. https://doi.org/10.1111/j.1756-8765.2008.01010.x

Kemp, S., Burt, C. D. B., & Furneaux, L. (2008). A test of the peak-end rule with extended autobiographical events. *Memory and Cognition*, *36*(1), 132–138. https://doi.org/10.3758/MC.36.1.132

Kemp, S., & Chen, Z. (2012). Overall hedonic evaluations and evaluation of specific moments from past relationships and high school days. *Journal of Happiness Studies*, *13*(6), 985–998. https://doi.org/10.1007/s10902-011-9302-6

Lew, T. F., & Vul, E. (2015). Ensemble clustering in visual working memory biases location memories and reduces the Weber noise of relative positions. *Journal of Vision*, *15*(4), 10. https://doi.org/10.1167/15.4.10

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Science*, *15*(8), 343–351. https://doi.org/10.1016/j.tics.2011.06.004

Lutz, N. D., Diekelmann, S., Hinse-Stern, P., Born, J., & Rauss, K. (2017). Sleep supports the slow abstraction of gist from visual perceptual memories. *Scientific Reports*, *7*(1), 1–9. https://doi.org/10.1038/srep42950

Murty, P. V., LaBar, K. S. ., & Adcock, R. A. (2016). Distinct medial temporal networks encode surprise during motivation by reward versus punishment. *Physiology & Behavior*, *176*(1), 100–106. https://doi.org/10.1016/j.nlm.2016.01.018.Distinct

Nosofsky, R. M. (1987). Attention and Learning Processes in the Identification and Categorization of Integral Stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87–108. https://doi.org/10.1037/0278-7393.13.1.87

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, *120*(2), 297–328. https://doi.org/10.1037/a0031541

Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning and Memory*, *24*(1), 65–69. https://doi.org/10.1101/lm.042978.116

Peng, S., Liu, C. H., Liu, W., & Yang, Z. (2022). Emotion matters: Face ensemble perception is affected by emotional states. *Psychonomic Bulletin and Review*, *29*(1), 116–122. https://doi.org/10.3758/s13423-021-01987-w

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*(2p1), 304–308. https://doi.org/10.1037/h0028558

Richards, B. A., Xia, F., Santoro, A., Husse, J., Woodin, M. A., Josselyn, S. A., & Frankland, P. W. (2014). Patterns across multiple memories are identified over time. *Nature Neuroscience*, *17*(7), 981–986. https://doi.org/10.1038/nn.3736

Rosch, Eleanor; Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Schmidt, S. R., & Schmidt, C. R. (2017). Revisiting von Restorff's early isolation effect. *Memory and Cognition*, *45*(2), 194–207. https://doi.org/10.3758/s13421-016-0651-6

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, *366*(6466). https://doi.org/10.1126/science.aau5141

Sherman, S. J., Sherman, J. W., Percy, E. J., & Soderberg, C. K. (2013). Stereotype development and formation. In *Oxford Handbook of Social Cognition*. https://doi.org/10.1093/oxfordhb/9780199730018.013.0027

Shigemune, Y., Tsukiura, T., Kambara, T., & Kawashima, R. (2014). Remembering with gains and losses: Effects of monetary reward and punishment on successful encoding activation of source memories. *Cerebral Cortex*, *24*(5), 1319–1331. https://doi.org/10.1093/cercor/bhs415

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436. http://doi.apa.org/.cfm?doi=10.1037/0278-7393.24.6.1411

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(1), 3–27. https://doi.org/10.1037/0278-7393.26.1.3

Tong, K., Dubé, C., Dubé, D., & Sekuler, · Robert. (2019). What makes a prototype a prototype? Averaging visual features in a sequence. *Attention, Perception, & Psychophysics*, 1–17. https://doi.org/10.3758/s13414-019-01697-5

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., … Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82. https://doi.org/10.1126/science.ll37786

Valentine, T. I. M., Darling, S., & Donnelly, M. (2004). *Why are average faces attractive? The effect of view and averageness on the attractiveness of female faces*. *11*(3), 482–487.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–129.

Ying, H. (2022). Attention modulates the ensemble coding of facial expressions. *Perception*, *51*(4), 276–285. https://doi.org/10.1177/03010066221079686

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *ELife*, *10*, 1–24. https://doi.org/10.7554/eLife.65588