Relatedness in memory and metamemory:

Benefits, costs, and beliefs

by

Xinyi Lu

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Psychology

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Chris Fiacconi Associate Professor Department of Psychology University of Guelph
Supervisors	Dr. Colin MacLeod Distinguished Professor Emeritus Department of Psychology
	Dr. Evan Risko Associate Professor Department of Psychology
Internal Members	Dr. Myra Fernandes Professor Department of Psychology
	Dr. Samuel Johnson Assistant Professor Department of Psychology
Internal-external Member	Dr. Michael Barnett-Cowan Associate Professor Department of Kinesiology & Health Sciences

Author’s Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This research was conducted at the University of Waterloo by Xinyi Lu under the supervision of Dr. Evan Risko and Dr. Colin MacLeod. An earlier version of Chapter 1 has been published in *Psychological Research* (Lu et al., 2023). This article was co-authored by myself, Mona Zhu, and Evan Risko. Chapter 2 is currently under review (Lu & Risko, under review) and was co-authored by myself and Evan Risko. Chapter 3 is currently under review (Lu et al., under review) and was co-authored by myself, Bhumika Bhandari, and Evan Risko.

Abstract

This dissertation examines the influence of semantic relatedness on both memory and metamemory. Related items tend to be better remembered than unrelated items in most memory tasks, and people are usually able to anticipate this in their memory predictions. In this dissertation, I report a novel case where inter-item relatedness produces a memory cost, specifically, in a location memory task. Despite this cost, participants predict that relatedness should be beneficial in this task, showing a misalignment between their beliefs and their performance. I then examine the mechanisms underlying the relatedness cost in location memory performance and what I refer to as the relatedness halo in metamemory. The latter phenomenon in particular is used to provide novel insights into the nature of metamemory beliefs and how they are updated in response to new information. I advance a theoretical framework for understanding beliefs as cue-dependent judgments that are constructed from multiple sources of retrieved information.

Acknowledgments

I am forever grateful to my advisors Evan Risko and Colin MacLeod for their guidance on this journey. I really could not have asked for better mentors. I hope to make a positive impact on people like you have made on me. Thank you.

To my labmates in arms, past and present. It has been a privilege to learn and grow with you. You deserve all your hopes and dreams and more.

To my husband, Hao. Your love and support are everything. Time for our next chapter.

Table of Contents

List of Figures

List of Tables

Introduction

Memory refers to our ability to encode and retrieve information, allowing us to learn from experience. Learning and remembering are fundamental skills that underlie much of human cognition. Metacognition refers to our ability to monitor, evaluate, and control our own cognitive processes (Nelson & Narens, 1990), thus metamemory is metacognition applied to memory. This dissertation investigates the effect of semantic relatedness on both objective memory performance and subjective metamemory predictions. The central conceit is that semantically related items, by virtue of their similarity to each other in semantic space, allow us to observe two sides of the same coin: memory facilitation based on shared semantic similarity/associations, and memory impairment based on similarity-based interference.

Although semantic memory facilitation is much more commonly observed, this dissertation focuses on a context where we observe overall semantic memory impairment. This work is structured into three chapters examining memory performance and/or prediction in a *location memory task* where to be remembered items in a display are either all related (*e.g., shirt, dress, shoes*) or unrelated (*e.g., ball, desk, pan*). In Chapter 1, I investigate the effect of relatedness on memory performance in this task which, unlike many other tasks, leads to a cost rather than a benefit. In Chapter 2, I investigate why participants nonetheless *believe* that relatedness is beneficial to memory even in this particular context where it is not, and I propose a *beliefs as constructive judgments* framework for understanding beliefs and their contributions to metacognitive judgments. Finally, in Chapter 3, I apply the framework in an investigation of understanding how participants' erroneous beliefs in this context change during the course of an experience—an experiment in which they can observe their own behaviour.

Relatedness in memory: the double-edged sword of semantic similarity

In most tasks, semantically related information tends to be better remembered than unrelated information. For example, a related list of words is usually recalled better compared to an unrelated list of words (Hunt & McDaniel, 1993; Lewis, 1971; Lu et al., 2022; Mandler, 1967; Puff, 1970), and related word pairs (e.g., table-chair) are usually remembered better compared to unrelated word pairs (e.g., table-water) in a cued recalled task (e.g., table-_____) (e.g., Castel et al., 2007; Epstein et al., 1975; Mueller et al., 2016). Under some circumstances, however, semantic relatedness can lead instead to memory costs; for example, in comparing the learning of related words vs. unrelated words, relatedness can lead to increased false recall of unpresented words (Guerin & Miller, 2008; Lu et al., 2022; Roediger & McDermott, 1995).

According to Kahana et al. (2022), semantic relatedness typifies the law of *similarity* in memory, a key principle around which memories are organized. Memory retrieval is often triggered by some similarity between the present and some past experience (Kahana, 2020; Surprenant et al., 2006). This idea has been formalized in models that conceive of retrieval as a search of memory in response to similar or matching cues at test (e.g., Kimball et al., 2007; Polyn et al., 2009; Raaijmakers & Shiffrin, 1981). Both the benefits and costs of semantic relatedness have been proposed to be a result of the same underlying principle, that is, relatedness promotes both generalization across and interference between semantically similar items (Kahana et al., 2022). Relatedness is something of a “double-edged sword” in memory; depending on the situation, it could lead to benefits and/or to costs, both within and across different tasks (Kahana et al., 2022). Thus, memory tends to be better for a related list of words than for an unrelated list of words as the shared semantic/category information that can be extracted from the related list serves as a powerful cue to access information during retrieval

(Mandler, 1967; Tulving & Pearlstone, 1966). The retrieval of one item on the list can also cue the retrieval of other related items on the list, as illustrated by the observation that individuals tend to make transitions among semantically related items in their free recall output (Bousfield et al., 1954; Romney et al., 1993), even when learning 'unrelated' lists without obvious semantic structure (Howard & Kahana, 2002b).

From a memory search perspective, candidates that match the given semantic cues are distinct compared to non-matching distractors, thus are more likely to be retrieved and to lead to a benefit in free recall. This benefit, however, goes hand in hand with decreased distinctiveness between similar items *within* a category (Hunt, 2013; Hunt & Einstein, 1981). This effect can be observed in participants' tendency to retrieve unpresented items that are nonetheless consistent with the semantic cues (e.g., believing that SLEEP was presented after studying BED, REST, AWAKE; Roediger & McDermott, 1995). Shared semantic similarity thus primes the true recall of targets (i.e., facilitation) but also leads to the false recall of associated distractors (i.e., interference).

In sum, semantic relatedness increases inter-item similarity and decreases individual item distinctiveness, which can result in either an overall benefit or a cost to memory depending on which factor is more influential based on the demands of the test conditions. In Chapter 1, I investigated the potential "double-edged" sword of semantic relatedness in the context of item location memory (Postma et al., 2008; Postma & De Haan, 1996). In contrast to a typical free recall task, in which participants are simply required to recall all items, a location memory task requires remembering that specific items belong to specific locations. Performance in this task hinges on distinguishing between these representations at retrieval (e.g., Hund & Plumert, 2003; Lu et al., 2024). Therefore, I predicted that increased inter-item semantic similarity would facilitate performance in the free recall task (i.e., item memory) but that this same similarity would be detrimental to performance in the location memory task (i.e., location memory).

Relatedness in metamemory: the role of beliefs

Theoretical accounts of metacognition, such as the dual-basis view (Koriat, 1997; Koriat et al., 2004) and analytic processing theory (Mueller & Dunlosky, 2017), emphasize a central role for *a priori* theories or beliefs about memory in our metacognitive judgments. In the case of semantic relatedness, participants tend to believe that semantically related information will be better remembered than unrelated information. They predict that related lists of words will be easier to learn and remember than unrelated lists (e.g., Chang & Brainerd, 2023; Koriat, 1997; Lu et al., 2022; Matvey et al., 2006; Mueller et al., 2013), and that related pairs of words will be easier to remember than unrelated pairs (Castel et al., 2007; Mueller et al., 2016). In both of these examples, participants are correctly anticipating a beneficial effect of relatedness on memory. These predictions are assumed to be (at least in part) based on pre-existing beliefs: When given a description of a hypothetical cued recall experiment, people estimate that they will remember more related than unrelated pairs (Mueller et al., 2013).

In Chapter 2, I solicited participants' predictions of memory in a location memory context, where relatedness leads to a true memory cost. This was done by presenting participants with a vignette describing a hypothetical location memory experiment like the ones reported in Chapter 1. Most participants predicted that they would remember more related than unrelated item locations, which was contrary to the true cost of semantic relatedness; only a minority of participants were able to anticipate a cost. I proposed a framework for understanding beliefs as constructive, cue-dependent judgments. In the framework, when participants' metamemory beliefs are solicited, they respond with a judgment that is constructed based on a retrieved subset of belief-relevant information. Belief-relevant information can be propositional/semantic knowledge (e.g., relatedness benefits memory; related items are easily connected; related items are less distinct) or specific episodes (e.g., related items were easier to remember in a previous task). Participants' inaccurate beliefs about the effect of relatedness can be accounted for by them being more likely to access the notion of memory facilitation based on the shared semantic similarity/associations than to access the notion of memory impairment based on similarity-based interference.

The framework was further applied in Chapter 3 to understand how belief reports are updated in response to new information. After making their initial predictions in response to the vignette and prompt (from Chapter 2), participants experienced the location memory task for themselves (from Chapter 1) before their predictions were solicited again. Their initial predictions tended to anticipate a benefit of relatedness, though their true task experiences tended to be a cost of relatedness. After first-hand experience with the location memory task, participants' final predictions showed a decrease in the predicted benefit of relatedness, although they did not reverse to predicting a cost of relatedness in the aggregate. Critically, changes in their belief reports followed the same direction as their task experience: Those who experienced a benefit shifted their relatedness predictions upward accordingly whereas those who experienced a cost of relatedness shifted theirs downward. While their final predictions were clearly based upon their objective memory performance to some extent, their previous predictions continued to influence their final predictions as well.

In the *beliefs as constructive judgments* framework, any report of a "belief" is a judgment that is constructed from multiple sources of retrieved information. Therefore, one's "beliefs" can differ depending on factors such as the elicitation context and the wording of the question, as these will influence retrieval cues and subsequently what information is available and used to construct a belief response. This view of beliefs positions them as judgments that are constructive and cue-dependent in the same way that item-by-item judgments of learning are (Koriat, 1997).

Chapter 1

Despite memory for semantically related items being improved over that for unrelated items in many cases, relatedness can also lead to memory costs. Here I examined how the semantic relatedness of words within a display influenced memory for their locations. Participants learned the locations of words inside grid displays; the words in a given display were either from a single category or were from different assorted categories. When a display containing words from a single category was compared to a scrambled display containing words from multiple categories, location memory performance was rendered worse in contrast to word recall performance which was significantly improved. These results suggest that semantically structured spaces can both help and harm memory within the context of a location memory task. I hypothesize that relatedness can improve memory performance by increasing the likelihood that matching candidates will be retrieved but can worsen performance that requires distinguishing between similar target representations.

Semantically related information tends to be better remembered than unrelated information. For example, a related set of words is usually recalled better compared to an unrelated set of words (Hunt & McDaniel, 1993; Lewis, 1971; Lu et al., 2022; Mandler, 1967; Puff, 1970). However, semantic relatedness can also, in certain tasks, lead to memory costs; for example, in comparing the learning of related words versus unrelated words, relatedness can lead to increased false recall of unpresented words (Guerin & Miller, 2008; Lu et al., 2022; Roediger & McDermott, 1995). Both the benefits and costs of semantic relatedness have been proposed to be a result of the same underlying principle, that is, it promotes both *generalization* across and *interference* between semantically similar items (Kahana et al., 2022). Semantic relatedness, therefore, might be something of a “double-edged sword” in memory; depending on the situation, it could lead to benefits and/or costs, both within and across different tasks (Kahana et al., 2022). In the present investigation I examine a prediction that arises henceforth; namely, how learning semantically related (vs. unrelated) items will influence memory for their locations.

According to Kahana et al. (2022), semantic relatedness typifies the law of similarity in memory, a key principle around which memories are organized. Memory retrieval is often triggered by some similarity between the present and some past experience (Kahana, 2020; Surprenant et al., 2006). This idea has been formalized in models that conceive of retrieval as a search of memory in response to similar or matching cues at test (e.g., Kimball et al., 2007; Raaijmakers & Shiffrin, 1981). In these accounts, memory for a related list of words tends to be better than an unrelated list of words as the categorical information that can be extracted from the former serves as a powerful cue to access information during retrieval (Mandler, 1967; Tulving & Pearlstone, 1966). The retrieval of one item on the list can also cue the retrieval of other related items on the list, as illustrated by the observation that individuals tend to make transitions among semantically related items in their free recall output (Bousfield et al., 1954; Romney et al., 1993), even when learning ‘unrelated’ lists without obvious semantic structure (Howard & Kahana, 2002b). From a memory search perspective, candidates that match the given semantic cues are distinct compared to non-matching distractors, thus are more likely to be retrieved leading to a benefit in free recall. This benefit, however, goes hand in hand with increased similarity (decreased distinctiveness) between items *within* a category (Hunt, 2013; Hunt & Einstein, 1981). This effect is readily observed in participants’ tendency to retrieve unpresented items that are nonetheless consistent with the semantic cues (e.g., believing that SLEEP was presented after studying BED, REST, AWAKE; Roediger & McDermott, 1995). That is, the generalizing effect of semantic relatedness facilitates the true recall of targets but also leads to the false recall of distractors (i.e., interference). In a free recall task, therefore, the effect of semantic relatedness is chiefly a boon to memory performance, but nonetheless comes with some costs to precision (i.e., an increase in false alarm rate).

Here I investigated the potential “double-edged” sword of semantic relatedness in the context of item location memory (Postma et al., 2008; Postma & De Haan, 1996). In contrast to a typical free recall task, in which participants are simply required to recall all items, a location memory task requires binding to-be-remembered items to specific locations (e.g., Hund & Plumert, 2003; Lu et al., 2024). Thus, while I might expect increased inter-item semantic similarity to provide a boost to free recall (i.e., item memory), this same inter-item similarity might increase interference across item-location pair bindings (i.e., location memory). To date, there has been surprisingly little empirical work investigating the effect of inter-item semantic relatedness on location memory. In the following sections, I first discuss previous investigations into the effects of semantic relatedness on a similar task (memory for serial order), before reviewing the handful of extant investigations into the effects of semantic relatedness on location memory.

Semantic Relatedness and Order Memory

One can draw parallels from the aforementioned location memory task to a serial recall or serial reconstruction task: a location memory task requires binding to-be-remembered items to specific spatial locations, while the latter tasks require binding to-be-remembered items to specific serial positions (Kowialiewski et al., 2023). While semantic similarity has been consistently found to facilitate item memory (i.e., increased items recalled), some studies have found that it is detrimental to order memory (i.e., increased rate of serial order errors, e.g., Saint-Aubin et al., 2005; Tse, 2009; Tse et al., 2011). Conceptualized within a “double-edged sword” framework of similarity and distinctiveness, semantic relatedness is thought to hinder order recall by increasing similarity among retrieval cues (i.e., interference). A strong version of this argument is that semantic similarity-based interference leads to a form of confusion known as an “interpretation problem”: when recalling an item, another item is erroneously recovered and recalled in the position of the targeted one (Tse, 2009). However, a number of studies have not found a cost of semantic relatedness on serial order (e.g., Saint-Aubin & Poirier, 1999; Guérard & Saint-Aubin, 2012; Neale & Tehan, 2007; Neath et al., 2022), although a recent meta-analysis by Ishiguro and Saito (2021) suggests that semantic similarity might have a small detrimental effect on order memory. Overall, the balance of evidence would suggest either a small cost or a null effect (but not a benefit) of semantic relatedness on order memory, in contrast to the robust finding of its benefit on item memory.

One possibility for the weak or perhaps even null effect of semantic similarity on order memory in previous studies is that these studies employ immediate serial recall tasks, as they are primarily concerned with questions of how item and order information are represented in short-term/working memory. In the traditional working memory model (Baddeley & Hitch, 1974), short-term memory is conceived as primarily relying on phonological representations (i.e., the phonological loop) while long-term memory relies on a semantic-based code. As this conception would suggest, acoustic similarity-based interference in short-term memory is well-established, while evidence for semantic similarity-based interference is equivocal (Baddeley, 1966a; 1966b). Conversely, there is a longstanding body of literature establishing robust semantic similarity-based interference in long-term memory tasks. For example, proactive interference refers to the phenomenon whereby, if a series of to be remembered items are presented for later recall, performance on the later items becomes progressively worse as interference from previous items is built up (Underwood, 1957). However, if the category of the to-be-remembered items is changed partway during the presentation of the list (for example, from flowers to games), then performance returns to near baseline level (i.e., release from proactive interference; Gardiner et al., 1972).

In sum, while evidence for a cost of semantic relatedness in short-term memory for order is weak to none, there is robust evidence that semantic relatedness leads to interference in memory tasks that operate at longer time frames. I have highlighted two such examples: semantic relatedness can increase the false recall of related critical lures (e.g., Roediger & McDermott, 1995) and lead to the buildup of category-specific interference in recall (Gardiner et al., 1972). In the current investigation, I examine the question of semantic similarity-based interference in an item location memory task, at typical delays for investigating the influences of long-term memory.

Semantic Relatedness and Location Memory

As noted above, there are few existing studies investigating the effect of inter-item semantic relatedness on location memory. Early research established that pre-existing notions of semantic categories can have a biasing effect on our memory for location and distance. For example, we tend to think of post offices as belonging in the same vicinity as other commercial establishments like banks, as opposed to educational institutions like schools (Hirtle & Mascolo, 1986). As a result, we might remember a post office and a bank as being closer to each other than they actually were (Hirtle & Mascolo, 1986). Further evidence of the semantic biasing of location memory comes from a study by Hund and Plumert (2003) that asked participants to learn the locations of items arranged in either semantically homogenous quadrants (containing related items) or scrambled quadrants (containing unrelated items). They found that participants tended to remember objects as being closer together when they were semantically related than when they were scrambled. The biasing effect of semantic relatedness in these above examples is consistent with the notion of category generalization (Kahana et al., 2022), that is, the representations of same-category items became more similar to each other but also more distinct from other-category items. In the context of location memory, relatedness appears to have the effect of pushing same-category items closer together in space (Hund & Plumert, 2003).

Given the above, we might reasonably predict that semantic relatedness will lead to either benefits and/or costs to location memory performance depending on task demands (i.e., whether performance will be helped by generalization or hindered by interference). One demonstration of a benefit of inter-item semantic relatedness comes from a recent series of studies by Tompary and Thompson-Schill (2021). Participants were asked to study items that were distributed across a few on-screen regions, such that each region contained mostly items within the same category (e.g., animals: *lion*, *giraffe*, *raccoon*), though items from other categories could be present as well. Tompary and Thompson-Schill (2021) found that location memory accuracy was greater for same-category items that were clustered in the same region, compared to same-category items located in a different region. For example, “raccoon” would be located with greater accuracy if it had originally been clustered with other animals, rather than if it had been within the household items cluster (e.g., *chair*, *bucket*, *bowl*). The authors proposed that since same-category items tended to be clustered near each other, participants were able to use the spatial cluster-category association and use the category label as a cue to retrieve the region associated with a particular item’s category (Tompary & Thompson-Schill, 2021). This led to a benefit for items that were within their category-themed cluster, but not for items outside the cluster. In other words, in this particular task, generalizing a category’s item locations to a given region would benefit the items that were actually located in that region (consistent with the category-region association), but not items that were inconsistent with that region.

In a similar series of studies, Lu and colleagues (2024) asked participants to learn the locations of items in a semantically partitioned display (i.e., an environment that was composed of four partitions, each containing items from a single category) as well as a scrambled display (where each partition contained an assortment of items from different categories). A semantically partitioned display, for example, could contain all clothing items (e.g., *shirt*, *dress*, *jacket*) in the top-left partition, and all office supplies (e.g., *stapler*, *notebook*, *ruler*) in the top-right partition, while a scrambled display would have an assortment of items scrambled across the four partitions. During test, participants were given each item one by one and asked to indicate its original location. Critically, half the test trials were cued trials, on which participants were explicitly told which partition the item was located; the rest were uncued trials, on which participants were not given this information. Overall, Lu and colleagues found that the semantically partitioned display led to a location memory benefit over the scrambled display. However, this benefit was eliminated on cued trials, when participants were explicitly told which partition the item was in. Lu and colleagues proposed that the benefit was driven by increased distinctiveness between semantic category-themed partitions, allowing participants to associate the general category of items with a given region (e.g., tools are all in the top-right section). On uncued trials, this association would benefit memory to the extent that participants could at least place them in that region (rather than somewhere else in the full display). On cued trials where this effect was controlled for, however, the benefit of semantic relatedness was entirely eliminated.

If semantic relatedness increases distinctiveness between category clusters in space, but also increases similarity-based interference across item-location pairs within a category cluster (Hunt, 2013; Hunt & Einstein, 1981), then I should expect to observe a memory cost in a task where the benefit of between-category distinctiveness is reduced or eliminated. The cued condition in Lu et al. (2024) provided a test of this hypothesis, as in this condition, knowing only the category-partition association would provide no benefit. However, Lu and colleagues found no semantic cost (or benefit) to location memory within a given partition. Thus, while Lu and colleagues found support for a benefit from increased distinctiveness between semantic partitions, they found no evidence of a cost for increased interference within a semantic partition. One possible reason for the null effect on within-partition memory obtained by Lu et al. (2024) is that, in the context of location memory, semantic relatedness does not increase similarity/interference across same-category items. Another possibility is that the null effect was due to the use of a partitioned display, wherein the partition-cueing manipulation used by Lu and colleagues may not have been entirely successful in eliminating the benefit of between-category distinctiveness.

In the present investigation, I sought to clarify the existence of a cost or a benefit of semantic relatedness to location memory in an unpartitioned display, comparing a semantically related items grid display (containing items from a single category) to a semantically unrelated items grid display (containing items from different categories). I expected that this kind of display would more closely match past item memory research that has demonstrated robust benefits of semantic organization, wherein items are often drawn from either a list composed of items from a single category or not (Lewis, 1971; Lu et al., 2022; Puff, 1970). This design allows the examination of the effect of semantic relatedness on location memory performance without the contribution of the partition-level (between-category) benefits.

Current Investigation

I report three pre-registered experiments (Experiment 1a: https://osf.io/ncfg6; Experiment 1b: https://osf.io/8nxb6; Experiment 2: https://osf.io/m4phu) comparing item location memory for a semantically homogenous item display to a scrambled display[1]. Display type (semantic vs. scrambled) was manipulated within-participants: in each experiment, participants completed two study-test blocks, one in which the display contained words all belonging to the same semantic category, and one in which the display contained a random array of words from different semantic categories. Within each block, participants were shown the target words in the display one by one in the study phase, and subsequently, performed an explicit location memory test where they were given each word one by one in random order and asked to indicate its location. In Experiments 1a and 1b, I included a final global free recall task as a measure of item memory.

[1] Chapter 1 uses the terminology of semantically homogenous vs. scrambled displays; subsequent chapters refer to related vs. unrelated item conditions. These are equivalent terms.

I expected that the semantically homogenous display would result in an improvement in item recall relative to the scrambled display, akin to how semantically blocked word lists lead to higher rates of recall than scrambled lists (e.g., Lewis, 1971; Lu et al., 2022; Puff, 1970). I predicted that the semantically homogenous display might result in a cost to location memory relative to the scrambled display, based on increased semantic similarity-based interference between the items (Saint-Aubin et al., 2005; Tse, 2009; Tse et al., 2011).

An additional aspect of interest in the current investigation was in the metacognitive consequences of interference. Previous studies have shown that participants experience semantically related word pairs as more fluent to process than unrelated word pairs (e.g., Undorf & Erdfelder, 2015), that is, they report that the related word pairs feel easier to process. However, as I anticipated semantic relatedness to lead to interference-based costs in the current location memory task, I wondered to what extent participants would be sensitive to this. Hence, I included a post-experiment exploratory probe, where I asked participants to rate the subjective difficulty of the semantic versus scrambled displays.

Experiments 1a and 1b

Methods

Experiments 1a and 1b were identical (1b was intended as a replication of 1a) and are described together.

Participants

I pre-registered a sample size of 90 participants for each experiment based on a power analysis estimating a small-medium effect size for within-subjects comparisons ($dz = .30$, $\alpha = .05$, two-tailed, within-subjects *t*-test) using G*Power (Faul et al., 2007). This effect size is a conservative estimate compared to previous, comparable experiments (Tompary & Thompson-Schill, 2021; Lu et al., 2024). I further anticipated this sample size to be a conservative estimate with respect to the generally more powerful mixed effects models that would be used (Quené & van den Bergh, 2004). After removing participants who did not meet the pre-registered criteria, this sample size was achieved in Experiment 1a (*N* = 90; 62 women, 24 men, 2 other, 2 undisclosed, *M* = 20.00 years, *SD* = 2.85), while two additional participants had to be removed in Experiment 1b for previously participating in similar experiments (*N* = 88; 62 women, 26 men, *M* = 20.21 years, *SD* = 3.69). Participants were undergraduate students at the University of Waterloo participating for course credit and provided informed consent.

Materials

Stimuli were words from six categories of household objects (see Appendix A). Each category contained ten possible exemplars. For each participant, one category was randomly chosen to populate the semantic (single category) grid; the other, scrambled (multiple category) grid was populated by randomly selecting two items from each of the remaining five categories. As each 5 x 5 grid had 25 clickable location squares, 10 of these contained items and 15 were empty squares. Item positions in each grid were pseudo-randomly assigned such that each row and column contained two items (cf. Siegel & Castel, 2018). The order of the semantic/scrambled grids was again randomized across participants. Examples of the two grids are shown in Figure 1 (note that participants never saw the grid with all items visible).

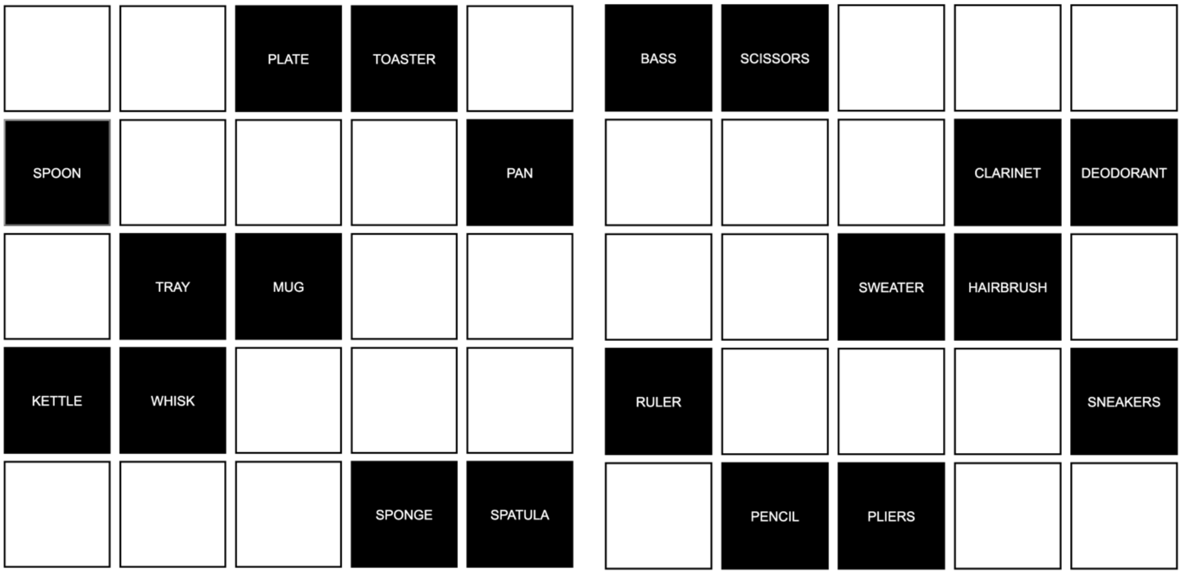


		PLATE	TOASTER	
SPOON				PAN
	TRAY	MUG		
KETTLE	WHISK			
			SPONGE	SPATULA

BASS	SCISSORS			
			CLARINET	DEODORANT
		SWEATER	HAIRBRUSH	
RULER				SNEAKERS
	PENCIL	PLIERS		

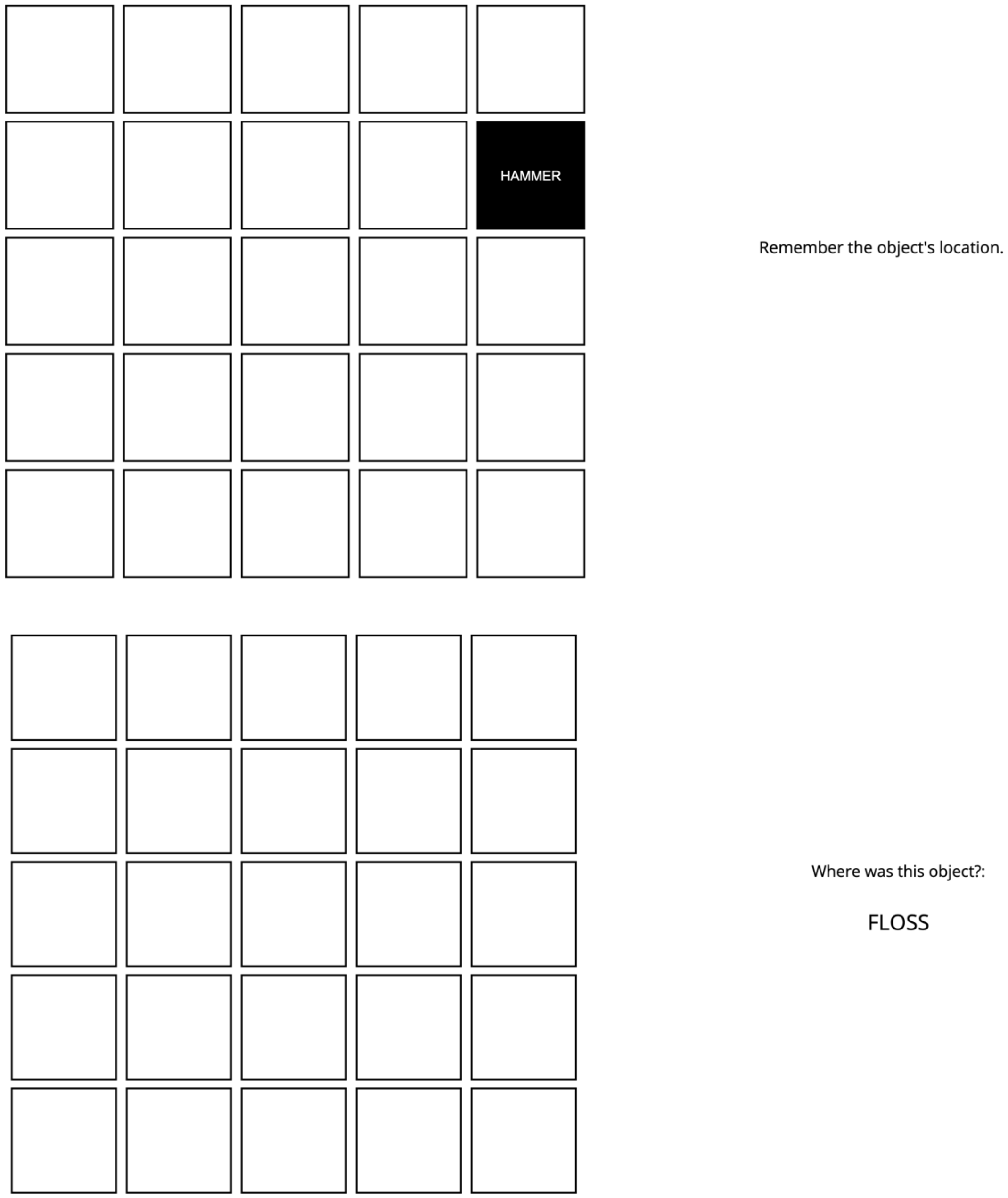
Figure 1. Semantic (left) and scrambled (right) item grids in Experiments 1a/1b.

Procedure

After being given a brief overview of the experiment, each participant performed two blocks, each containing the encoding task, the location memory task, and a difficulty probe. These tasks are described below.

Encoding Task. On each trial, only a single item in the grid was visible to the participant (see Figure 2). The visible item was presented for 5000 ms in white text on a black square. There were 10 trials for a single grid (one for each item, presented in random order) and the intertrial interval was 400 ms.

Location Memory Task. On each trial, participants were presented with the empty grid from the encoding task to the left. The target item was shown on the right. They were instructed to click on the square that corresponded to the original location of the target item. There were 10 trials for a single grid (one for each item, presented in random order) and no feedback was given.



HAMMER

Remember the object's location.

Where was this object?:

FLOSS

Figure 2. Encoding task (top) and location memory task (bottom).

Difficulty. Upon completion of the location memory task, participants were asked to answer, “How difficult was the location memory task you just did?” on a 7-point scale anchored by 1: Very Easy and 7: Very Difficult.

After participants completed both the semantic and scrambled blocks, they were asked to recall as many items as they could from both grids by typing them into an on-screen text field (i.e., global free recall task). Upon completion of this task, they were probed on (1) whether they noticed anything about the two grids; (2) what strategies they used to remember the locations of items. Participants were then asked to indicate which of the two grids they found to be easier. Finally, participants were asked to provide their age and gender and complete an attention and effort check questionnaire.

Results

Data from 16 participants in Experiment 1a and 40 participants in Experiment 1b were not analyzed according to the pre-registered criteria; they self-reported that they were not paying attention or did not give effort during the task in the post-study questionnaire. An additional 2 participants in Experiment 1b were excluded due to previously participating in related experiments. After all exclusions, this left a sample size of $N = 90$ in Experiment 1a and $N = 88$ in Experiment 1b. The analyses reported below were pre-registered unless stated otherwise. All analyses were conducted using R (R Core Team, 2019). Mixed-effect regressions were conducted using the *lme4* package (Bates et al., 2015). Categorical predictors (e.g., semantic vs. scrambled displays) were coded in the models using sum-contrasts. For the random effects structure, I began with a model containing by-participant and by-stimuli random intercepts; by-participant and by-stimuli random slopes for the effect of semantic/scrambled display were included only when doing so significantly improved the fit of the model (Bates et al., 2018; Matuschek et al., 2017). In cases where the initial model was singular (indicating possible overfitting), I removed either the by-participant or by-stimuli random intercept to reduce model complexity. Unless otherwise specified, all linear and logistic mixed-effects models were run using the *bobyqa* optimizer. Given that degrees of freedom can be difficult to estimate accurately in mixed-effects models (Bates et al., 2015), approximated *p*-values using Wald *z*-statistics are provided, considering the relatively large number of observations in the current study. This was done using the *sjPlot* package (Lüdecke, 2022). All data and analysis code are available at https://osf.io/8xf6a/.

Location Memory

In order to examine whether the semantic grid influenced location memory performance, I looked at (1) absolute location accuracy, defined by whether participants chose the correct target square or not; (2) Euclidean distance from the chosen location to the correct location on all trials.

Location Accuracy. A mixed-effects logistic regression revealed that there was a significant main effect of condition in Experiment 1a, such that location memory accuracy was lower in the semantic grid than the scrambled grid, *b* = -0.18, 95% CI [-0.28, -0.08], *z* = 3.61, *p* < .001. This effect was not significant in Experiment 1b, *b* = -0.09, 95% CI [-0.21, 0.03], *z* = 1.37, *p* = .170. Figure 3A shows the effect of semantic grid on location accuracy.

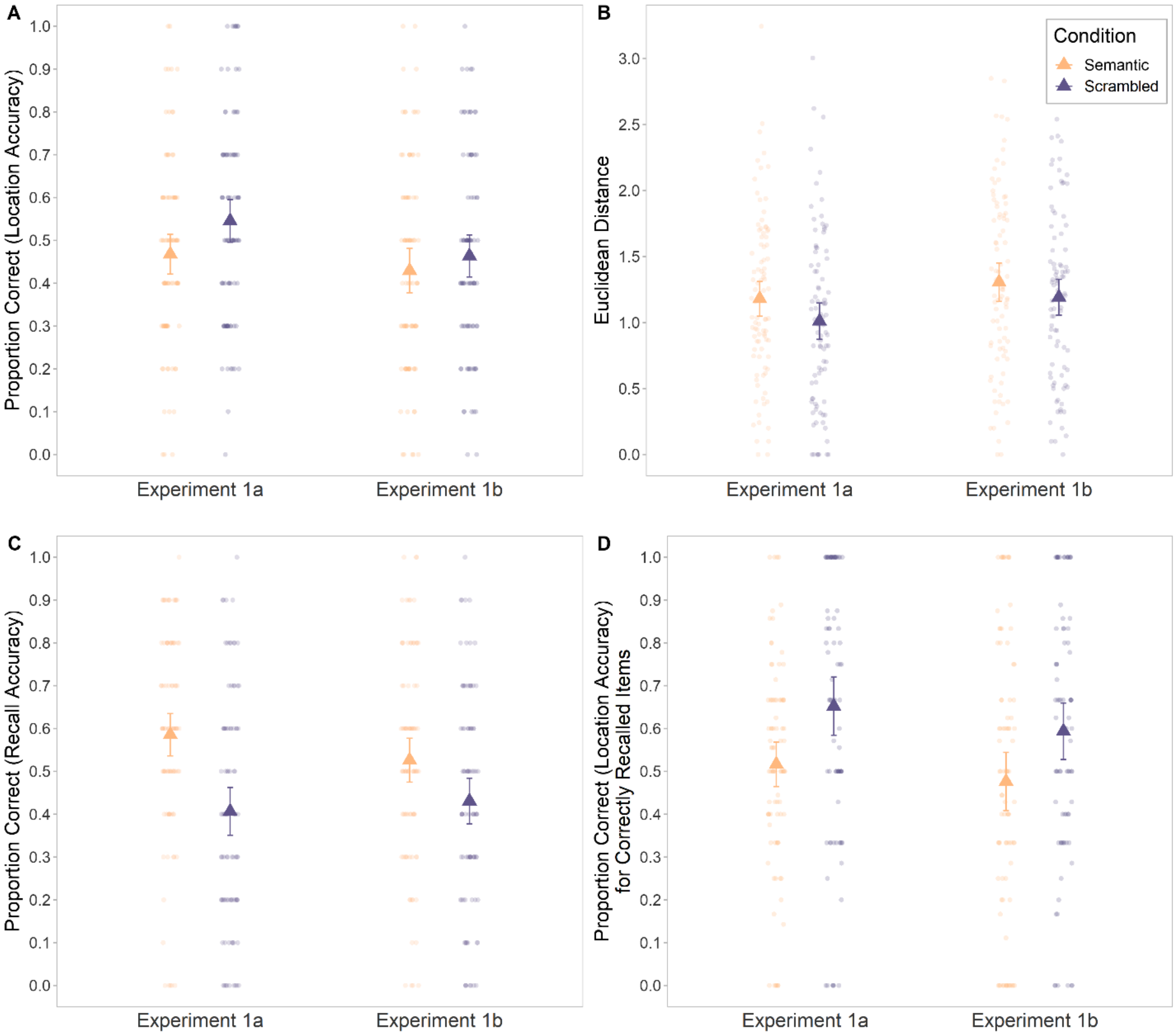


Figure 3. Mean (A) location accuracy, (B) Euclidean distance, (C) item recall accuracy, and (D) location accuracy for correctly recalled items only by semantic/scrambled display in Experiment 1a and 1b. Error bars represent 95% confidence intervals. Circles in the background represent each participant's average performance per condition.

Euclidean Distance. A mixed-effects linear regression revealed that there was a significant main effect of condition in Experiment 1a, such that mean Euclidean distance between selected and target locations was larger in the semantic grid than the scrambled grid, b = 0.09, 95% CI [0.03, 0.15], t = 2.68, p = .004. This effect was not significant in Experiment 1b, b

= 0.06, 95% CI [-0.02, 0.14], t = 1.45, p = .141. Figure 3B shows the effect of semantic grid on Euclidean distance.

Item Memory

Item Recall Accuracy. The semantic grid significantly improved item recall performance in both experiments, Exp 1a: b = 0.44, 95% CI [0.29, 0.60], z = 5.29, p < .001; Exp 1b: b = 0.19, 95% CI [0.03, 0.36], z = 2.27, p = .023. Figure 3C shows the effect of semantic grid on recall accuracy.

Intrusion Rate. The overall recall intrusion rate was very low, Exp 1a: 2.8%; Exp 1b: 3.4%. I performed a post-hoc classification of intrusions into the following categories: (1) related or similar to a word in the scrambled display; (2) related or similar to a word in the semantic display; (3) unrelated intrusion. The raw intrusion counts per category were: Exp 1a: scrambled: 11, semantic: 9, other: 6; Exp 1a: scrambled: 11, semantic: 16, other: 3.

Location Memory Controlling for Item Memory

In line with my predictions, the semantic grid improved item memory performance relative to the scrambled grid. Therefore, in order to assess the effect of the semantic grid on location memory performance alone, I attempted to account for the semantic benefit to item memory by statistically controlling for item memory performance. I entered both item recall accuracy and condition (semantic vs. scrambled) into a mixed-effects linear regression predicting location memory accuracy. There was a main effect of item recall on location memory performance, Exp 1a (exploratory): b = 0.77, 95% CI [0.55, 0.99], z = 6.92, p < .001; Exp 1b (pre-registered): b = 0.78, 95% CI [0.56, 1.01], z = 6.96, p < .001, suggesting that location memory accuracy was higher for items that were correctly recalled. Importantly, when controlling for item recall, the semantic grid was associated with reduced location memory performance relative to the scrambled grid, Exp 1a (exploratory): *b* = -0.26, 95% CI [-0.36, -0.15], *z* = 4.83, *p* < .001; this result was not significant in Exp 1b (pre-registered): *b* = -0.12, 95% CI [-0.25, 0.01], *z* = 1.83, *p* = .067.

Item Location Substitution Rate

In an exploratory analysis, I examined whether display type influenced how likely participants were to choose a location square containing another (incorrect) item when making an error. A mixed-effects linear regression showed that there was no difference between the semantic compared to scrambled grids, Exp 1a: *b* = 0.04, 95% CI [-0.13, 0.20], *z* = 0.43, *p* = .667; Exp 1b: *b* = -0.01, 95% CI [-0.14, 0.12], *z* = 0.15, *p* = .885.

Subjective Difficulty

I examined subjective difficulty in a series of exploratory analyses. I first conducted a mixed-effects linear regression on perceived task difficulty. As there was a significant effect of block, I report the model with condition and block, as well their interaction, as predictors. Participants reported that the semantic grid was more difficult than the scrambled grid, Exp 1a: *b* = 0.09, 95% CI [0.06, 0.12], *z* = 6.23, *p* < .001, Exp 1b: *b* = 0.07, 95% CI [0.04, 0.10], *z* = 4.66, *p* < .001. They also found the second block to be more difficult than the first, Exp 1a: *b* = -0.14, 95% CI [-0.17, -0.11], *z* = 9.76, *p* < .001, Exp 1b: *b* = -0.09, 95% CI [-0.12, -0.06], *z* = 6.10, *p* < .001. The interaction term was not significant, Exp 1a: *b* = -0.22, 95% CI [-0.47, 0.02], *z* = 1.77, *p* = .076, Exp 1b: *b* = -0.09, 95% CI [-0.30, 0.12], *z* = 0.83, *p* = .407.

I also conducted a linear regression using participants' semantic vs. scrambled performance differences to predict their subjective difficulty differences. I found that their semantic-scrambled performance differences negatively predicted their semantic-scrambled difficulty rating differences, Exp 1a: *b* = 2.52, 95% CI [1.55, 3.50], *t* = 5.14, *p* < .001, Exp 1b: *b*

= 2.81, 95% CI [2.02, 3.59], *t* = 7.01, *p* < .001. That is, participants who experienced a larger semantic performance cost also rated the semantic display as more difficult.

When explicitly asked to indicate which of the two displays was easier, participants were evenly split between the scrambled grid and the semantic grid, Exp 1a: 43 vs. 47, Exp 1b: 43 vs. 46 (one participant did not record a response). However, I found that these preferences were associated with differences in semantic-scrambled display performance, Exp 1a: *t*(87.96) = 5.40, *p* < .001, Exp 1b: *t*(83.64) = 5.65, *p* < .001. That is, participants who thought the scrambled display was easier also experienced a larger semantic performance cost (Mean semantic-scrambled performance difference: Exp 1a = -0.20, Exp 1b = -0.18), compared to participants who preferred the semantic display (Mean semantic-scrambled performance difference: Exp 1a = 0.03, Exp 1b = 0.10).

These results suggested that participants had a bias towards the semantic display, such that experience with a larger semantic performance cost was needed to shift participants' preferences away from the semantic display. To confirm this, I conducted a logistic regression to identify the performance point at which participants were indifferent between the two displays. An unbiased participant who had zero performance difference between the two displays would therefore have a 50% chance of selecting either display. The semantic-scrambled performance difference was significantly predictive of display preference, Exp 1a: *b* = 5.67, 95% CI [3.23, 8.60], *z* = 4.17, *p* < .001; Exp 1b: *b* = 5.48, 95% CI [3.18, 8.24], *z* = 4.28, *p* < .001. Importantly, the indifference point was associated with a negative semantic-scrambled performance difference (Exp 1a = -0.10, Exp 1b = -0.05), indicated that participants were biased towards selecting the semantic display as easier.

Discussion

Across two experiments, I found that participants performed worse (Experiment 1a) or the same (Experiment 1b) in a location memory task when presented with a display consisting of items from a single category compared to a scrambled display (consisting of items from different categories). Specifically, I found that (1) overall location accuracy either decreased in the semantic display or remained the same; (2) the average distance between the chosen location and the target location was either greater in the semantic display or remained the same. On the other hand, I found a significant benefit of the semantic display on item recall performance. When I statistically controlled for the benefit of item recall on location memory performance, I found evidence that the semantic display was associated with a reduction in location memory relative to the scrambled display; this result was significant in Experiment 1a but not in Experiment 1b. Altogether, these results provide evidence that a semantically related item display facilitated participants' ability to recall the items, but either reduced (Experiment 1a) or did not influence (Experiment 1b) their ability to correctly recall the item locations. That said, the results were less clear in Experiment 1b, and this was true for both item location memory and item memory. This might have reflected lower overall data quality, as suggested by the fact that many more participants in the sample had to be excluded due to self-reported inattention, and this may have inflated Type II error rate.

In an exploratory analysis, I compared item location substitution rate on participants' error trials for the two displays. This was motivated by the possibility that increased interference in the semantically related item display would manifest in increased confusability between semantically related items—that is, an increased likelihood to confuse item locations with each other in the semantic display when making an error. However, the two displays did not differ in item location substitution rate, suggesting that participants were not more likely to confuse item locations with each other in the semantic display when making an error. If semantic relatedness indeed increased interference between same-category representations (as was clear in E1a), then it did not appear to do so via whole item substitutions/confusions.

Additional notable findings from the exploratory analyses of Experiments 1a and 1b were that (1) participants reported that the semantic display was more difficult than the scrambled display; (2) participants' objective semantic-scrambled performance predicted their subjective reports, such that participants who experienced a larger semantic performance cost also rated the semantic display as more difficult; and (3) in spite of the aforementioned two points, participants actually exhibited a bias *towards* the semantic display, such that experience with a larger semantic performance cost was needed to shift participants' perception that the semantic display was easier. That is, despite firsthand experience that the semantic display was associated with poorer performance, they nonetheless selected it as the easier display in higher proportions than would be expected from their objective performance. This could suggest that participants might have had some initial notion that the related display is easier, which was then modified after their experience performing the actual task.

Experiment 2

In Experiment 1a, the results suggested that there was a cost in location memory associated with the semantically homogenous display. A similar trend was found in Experiment 1b, though not statistically significant. The goal of Experiment 2 was to replicate the results of Experiment 1a/b with an improved experimental manipulation and a larger sample size.

In Experiment 2, I selected 11 categories from the Van Overschelde et al. (2004) norms as stimuli. I selected categories that were high in category potency (a measure of how many items participants could generate to each category label) so that participants would be likely to recognize each item as belonging with the given category. A full list of the items can be found in Appendix A. Another change made in Experiment 2 was to use 10 words that were all from 10 different categories in the scrambled display, rather than 2 words from 5 different categories as in Experiment 1a/b. That is, the scrambled display was a 'truer' scrambled display in the sense that none of the words belonged to the same category, thus serving as a stronger manipulation relative to the semantically homogenous display. Furthermore, I added a distractor task between the encoding task and the location memory task so as to eliminate possible contributions from short-term memory. Finally, I pre-registered the item location substitution rate analysis that had been exploratory in Experiment 1a/b.

In terms of subjective difficulty, I also aimed to conceptually replicate the finding that participants might have a bias towards preferring the semantically related display, in spite of objective performance evidence to the contrary. Rather than asking about difficulty, I asked participants which of the two types of displays they would prefer to learn for an upcoming task. I predicted that while their location accuracy performance would be superior in the scrambled display, participants would prefer the semantic display after controlling for objective performance.

Methods

Participants

I pre-registered a sample size of 130 to achieve 80% power based on a power analysis using Superpower (Lakens & Caldwell, 2021) using the estimates obtained from Experiments 1a and 1b (M_1= 0.45, M_2= 0.51, SD = 0.23, r = 0.45). After exclusions, I analyzed complete data from 129 participants (65 women, 62 men, 1 other, 1 unknown, *M* = 40.25 years, *SD* = 13.35). Participants were recruited from Prolific and were paid 1.90 GPB.

Materials

Stimuli were words from eleven categories adapted from Van Overschelde et al. (2004) and can be found Appendix A. Each category contained ten possible exemplars. For each participant, one category was randomly chosen to populate the semantic (single category) grid; the other, scrambled (multiple category) grid was populated by randomly selecting one items from each of the remaining ten categories. As in Experiments 1a and 1b, item positions in each grid were pseudo-randomly assigned such that each row and column contained two items (cf. Siegel & Castel, 2018), and the order of the semantic/scrambled grids was again randomized across participants.

Procedure

The main experimental procedure was identical to Experiment 1a and 1b except with the addition of a distractor task between the encoding and location memory tasks in each block. Each participant performed two blocks, with each block containing the encoding task, a distractor task, and the location memory task. On each two-minute distractor task, participants were given a series of simple arithmetic statements (e.g., 1+1=2) and had to indicate whether each statement was True or False. Each statement was presented for up to 10 seconds, or until participants made a response. After participants completed both the semantic and scrambled blocks, they were asked to indicate (1) which of the two types of displays they would prefer to learn for an upcoming task, and (2) what strategies they used to remember the locations of items. Finally, participants were asked to provide their age and gender and complete an attention and effort check questionnaire.

Results

Data from 4 participants were not analyzed according to the exclusion criteria set in the pre-registration; they either did not adequately complete the arithmetic distractor task, and/or self-reported that they were not paying attention or did not give effort during the task, leaving a final sample size of *N* = 129. Data and analysis code are available at https://osf.io/8xf6a/. Mixed-effects models were used to examine each variable of interest (e.g., logistic for location accuracy and linear for distance) with display (semantic/scrambled) as a fixed effect; random effects structure was determined via model comparison as previously described in Experiment 1.

Location Memory

In order to examine whether the semantic grid influenced location memory performance, I examined (1) absolute location accuracy, defined by whether participants chose the correct target square or not; and (2) Euclidean distance from the chosen location to the correct location on all trials.

Location Accuracy. A mixed-effects logistic regression revealed that there was a significant main effect of condition, such that location memory accuracy was lower in the semantic grid than the scrambled grid, *b* = -0.12, 95% CI [-0.23, -0.01], *z* = 2.15, *p* = .031. Figure 4 shows the effect of display type on location accuracy.

Euclidean Distance. A mixed-effects linear regression revealed no significant main effect of condition, *b* = 0.05, 95% CI [-0.02, 0.11], *t* = 1.39, *p* = .166.

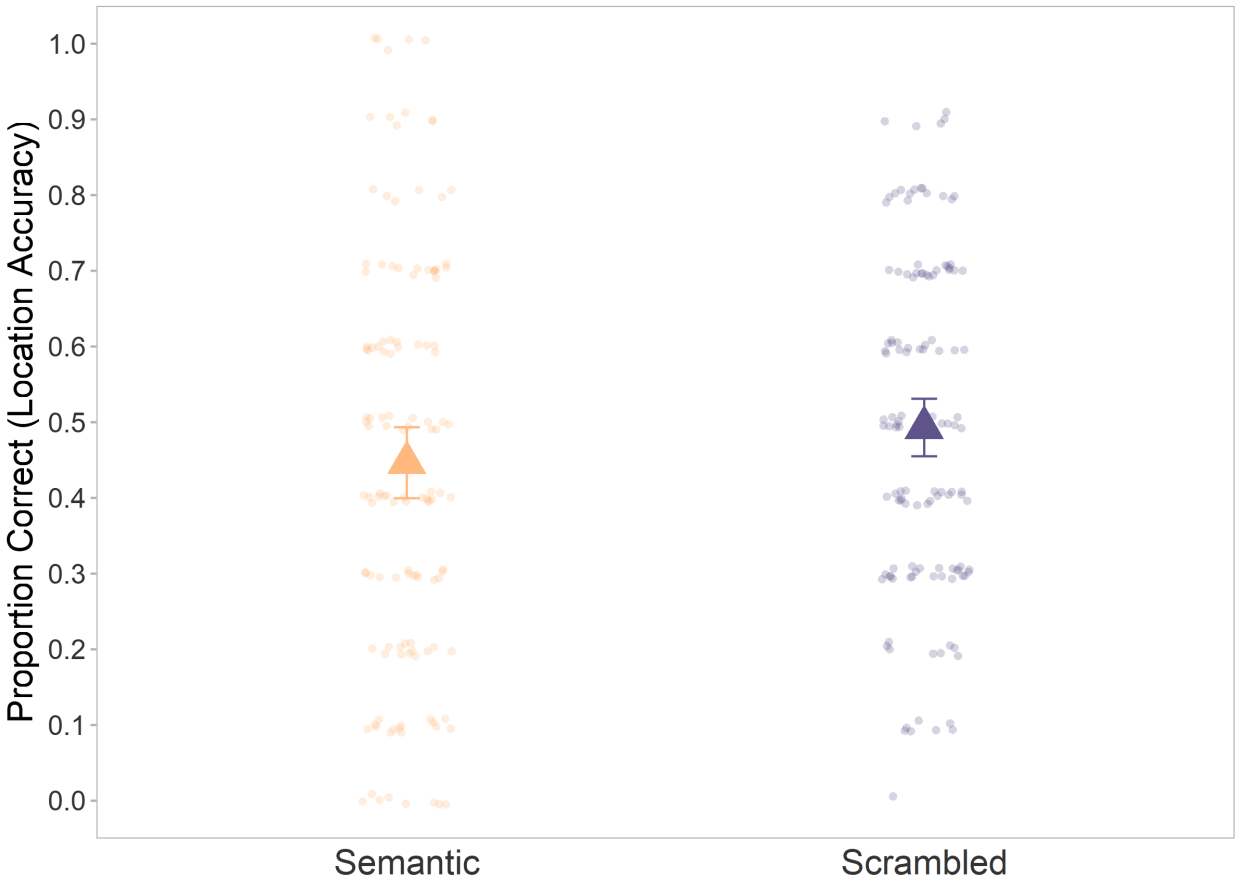


Figure 4. Mean location accuracy by semantic/scrambled display in Experiment 2. Error bars represent 95% confidence intervals. Circles in the background represent each participant's average performance per condition.

Item Location Substitution Rate

A mixed-effects logistic regression showed that participants were more likely to substitute another item location when making an error in the semantic compared to scrambled grid, b = 0.11, 95% CI [0.01, 0.22], z = 1.98, p = .048.

Participant Choice

When explicitly asked to choose one of the two display types for a future task, more participants preferred the semantic display than the scrambled display (76 vs. 53). This difference was significantly different from chance, χ2(1) = 4.10, *p* = .043. I found that these two groups of participants (semantic preference group vs. scrambled preference group) were associated with differences in semantic-scrambled display performance, *t*(121.83) = 7.12, *p* < .001. That is, participants who preferred the scrambled display had experienced a larger semantic performance cost (Mean semantic-scrambled performance difference = -0.21), compared to participants who preferred the semantic display (Mean semantic-scrambled performance difference = 0.07). I conducted a logistic regression to identify the performance point of indifference between the two displays. The semantic-scrambled performance difference was significantly predictive of display preference, *b* = 5.39, 95% CI [3.50, 7.61], *z* = 5.17, *p* < .001. The indifference point was predicted by a semantic-scrambled performance difference of -0.14, suggesting that participants by and large preferred the semantic display, but experience with a large semantic cost was able to shift this preference.

Discussion

In Experiment 2, I replicated the finding that participants performed worse in a location memory task when presented with a display consisting of items from a single category compared to a scrambled display consisting of items from different categories. Specifically, I found that overall location accuracy was significantly decreased in the semantic display; the average distance between the chosen location and the target location did not differ between the two displays. Together with Experiment 1a/b, these results provide evidence that a semantically related item display reduced the ability to correctly recall the item locations. These results are consistent with the notion that semantic relatedness between items on a list renders them more similar to each other and less distinct, thereby increasing interference at retrieval. While I did find that participants were more likely to substitute another item location when making an error in the semantic compared to scrambled grid, this effect was not robust across experiments. In terms of participants' preferences, I found that more participants chose the semantic display than the scrambled display when explicitly asked to choose one of the two display types for a future task. This finding is surprising given that participants performed significantly *worse* in this display. While participants by and large preferred the semantic display, experience with a large semantic cost was apparently able to shift this preference, as participants who selected the scrambled display tended to be those who had experienced a larger semantic performance cost. These results suggest that participants might have a bias towards the semantic display, and their preferences were only partially rooted in objective performance considerations.

Experiment 3

In Experiment 2, I replicated the finding that participants performed worse in a location memory task when presented with a display consisting of items from a single category compared to a scrambled display consisting of items from different categories. Together with Experiment 1a/b, these results provide evidence that a semantically related item display reduced the ability to correctly recall the item locations. However, the evidence for a whole item confusion-based account was not robust: while I did find a significant effect in Experiment 2, such that participants were more likely to substitute the location of one item for another in the semantic display, the size of the effect was small, and this effect was not found in Experiment 1a and 1b. The goal of Experiment 3 was to investigate an alternative hypothesis for the cost of semantic relatedness, that participants may have put in less effort into studying the item locations in the semantically homogenous display as they believed this type of display to be easier. Previous studies have shown that participants harbor beliefs that related items are easier to remember; for example, when given a description of a hypothetical memory experiment, participants predict that they will remember more related word pairs (e.g., cow-milk) compared to unrelated word pairs (e.g., fish-pen; Mueller et al., 2013). Given a self-paced study task, participants also tend to spend less time studying related word pairs compared to unrelated word pairs, though they still demonstrate better memory for the related than unrelated pairs (e.g., Castel et al., 2007; Mueller et al., 2016). In Experiment 1a and 1b, I found that participants exhibited a bias towards the semantic display: despite having just experienced that this display was associated with poorer performance, they nonetheless selected it as the easier display in higher proportions than would be expected from their objective performance. In Experiment 2, I also found that more participants selected the semantic display than the scrambled display when explicitly asked to choose one of the two display types for a future task, again, despite firsthand experience that the semantic display was associated with poorer performance. These results could suggest that participants might have had some initial notion that the semantically related items display was easier to learn, which could have led them to put less effort into studying these items, and/or putting in more effort into studying the scrambled items to compensate. To investigate this "metacognitive loafing" hypothesis, Experiment 3 used a self-paced version of the location memory learning task, where participants were allowed to freely vary the amount of time they spent studying the location of each item as it was presented. I predicted that if the metacognitive loafing hypothesis was correct, then participants would spend less time studying the items in the semantic display compared to the scrambled display, and that the cost to location memory associated with the semantic display would be mediated by this reduced study time.

Methods

Participants

I pre-registered a sample size of 220 to achieve 80% power based on a power analysis using Superpower (Lakens & Caldwell, 2021) using the estimates obtained from Experiment 2 (M_1= 0.44, M_2= 0.49, *SD* = 0.24, *r* = 0.42). After exclusions, I analyzed complete data from 220 participants (88 women, 121 men, 3 other, 8 unknown, *M* = 38.05 years, *SD* = 12.74). Participants were recruited from Prolific and were paid 1.90 GPB.

Materials

Stimuli were words from eleven categories of household items and can be found in Appendix A. Each category contained ten possible exemplars. For each participant, one category was randomly chosen to populate the semantic (single category) grid; the other, scrambled (multiple category) grid was populated by randomly selecting one items from each of the remaining ten categories. As in previous experiments, item positions in each grid were pseudo-randomly assigned such that each row and column contained two items (cf. Siegel & Castel, 2018), and the order of the semantic/scrambled grids was again randomized across participants.

Procedure

The main experimental procedure was identical to Experiment 2 except that the encoding task was self-paced. On each encoding trial, participants were presented with a single visible item in a grid. Participants were told that they were to remember each item’s location as it was presented, and that they could control how much time they spent study each item. Participants were instructed to click on a button labeled ‘Continue’ or anywhere on the grid to advance to the next trial when they felt they had completed studying an item. There were 10 trials for a single grid (one for each item, presented in random order) and the intertrial interval was 400 ms.

After completing both the semantic and scrambled blocks, participants were asked to provide their age and gender and complete an attention and effort check questionnaire.

Results

Data from 8 participants were not analyzed according to the exclusion criteria set in the pre-registration; they either did not adequately complete the arithmetic distractor task, and/or self-reported that they were not paying attention or did not give effort during the task. Due to an error in condition assignment, an additional 35 participants that were collected in excess of the stopping rule were excluded, leaving the final pre-registered sample size of $N = 220$. Data and analysis code are available at https://osf.io/8xf6a/. Mixed-effects models were used to examine each variable of interest (i.e., logistic for location accuracy and linear for distance) with display (semantic/scrambled) as a fixed effect. As outlined in the pre-registration, I compared these models to the models that included the logarithm of study time as a fixed factor. Random effects structures were determined via model comparison.

Study Time

Study Time. A mixed-effects linear regression revealed that there was a significant main effect of condition, such that study times were shorter in the semantic grid ($M = 5300$ ms) than the scrambled grid ($M = 5627$ ms), $b = -0.04$, 95% CI [-0.07, -0.00], $t = 2.06$, $p = .039$.

Location Memory

Location Accuracy. A mixed-effects logistic regression revealed that there was a significant main effect of condition, such that location memory accuracy was lower in the semantic grid than the scrambled grid, $b = -0.10$, 95% CI [-0.19, -0.01], $z = 2.29$, $p = .022$. When log(study time) was included in the model, longer study times were associated with increased location accuracy, $b = 0.53$, 95% CI [0.42, 0.64], $z = 9.59$, $p < .001$, while the effect of condition was marginally significant, *b* = -0.08, 95% CI [-0.16, 0.00], *z* = 1.93, *p* = .054. Figure 5A shows the effect of semantic grid on location accuracy.

Euclidean Distance. A mixed-effects linear regression revealed a significant main effect of condition, such that participants chose further away from the target location in the semantic grid than the scrambled grid*,* *b* = 0.05, 95% CI [0.01, 0.10], *t* = 2.24, *p* = .025. When log(study time) was included in the model, longer study times were associated with shorter distances from the target, *b* = -0.28, 95% CI [-0.34, -0.23], *z* = 9.78, *p* < .001, and the effect of condition was marginally significant, *b* = 0.04, 95% CI [0.00, 0.09], *z* = 1.91, *p* = .056. Figure 5B shows the effect of semantic grid on Euclidean distance.

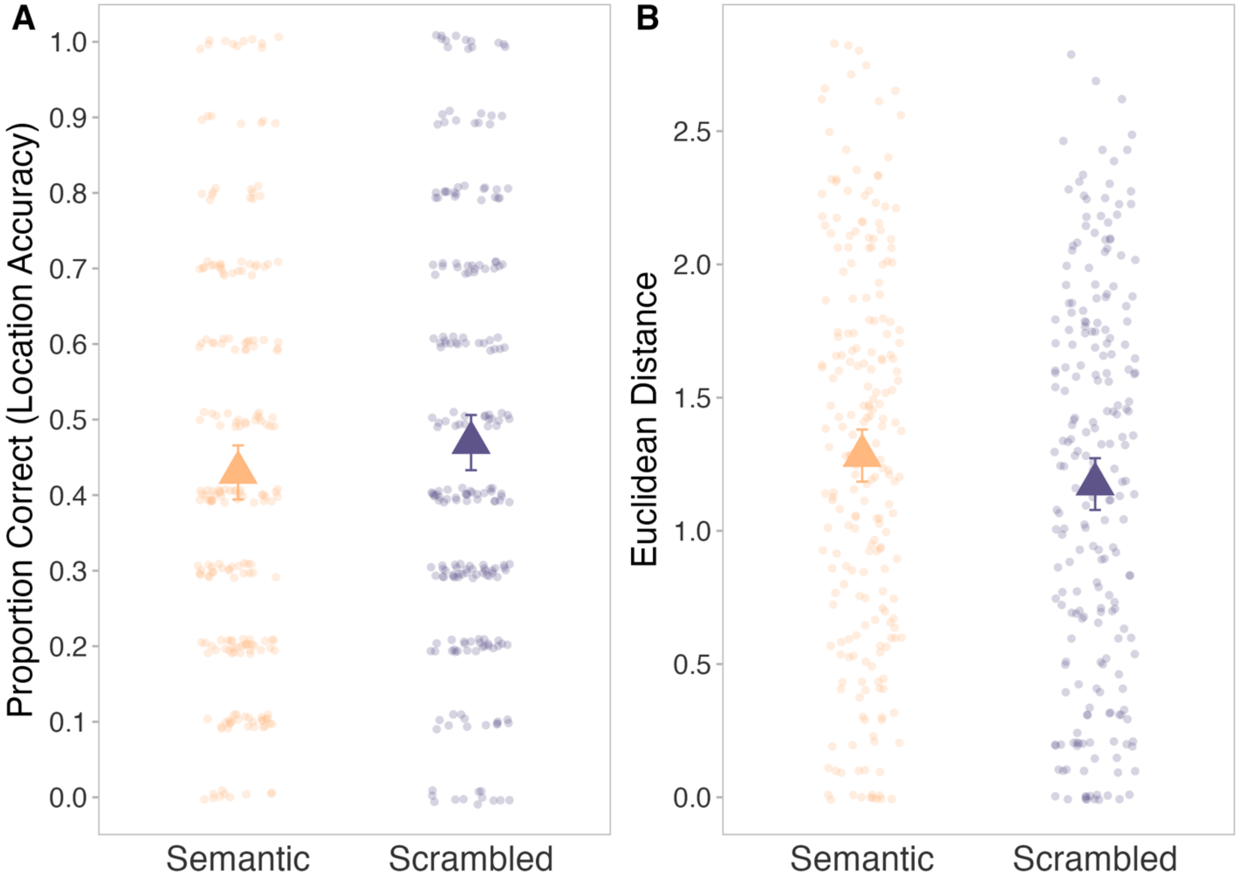


Figure 5. Mean (A) location accuracy, and (B) Euclidean distance by semantic/scrambled display in Experiment 3. Error bars represent 95% confidence intervals. Circles in the background represent each participant's average performance per condition.

Study Time Mediation Analysis

I performed a Bayesian mediation analysis using the *bmlm* package in R (Vuorre, 2016) with default priors (Normal(0, 1000) for regression coefficients, and Cauchy(0,50) for subject-level standard deviations). Display type (relatedness) was set as the predictor variable, log(study time) as the mediating variable, and location accuracy as the outcome variable. 95% credible intervals were computed from the posterior distribution of the model parameters using Markov chain Monte Carlo (MCMC) procedures with 10,000 iterations. Mirroring the earlier analyses, the semantic display type was predictive of shorter study times, *b* = -0.07, 95% CI [-0.14, -0.00], and increased study time had a positive effect on location accuracy, *b* = 0.26, 95% CI [0.11, 0.41]. However, the effect of display type on location accuracy was not mediated by study time: the mediated effect was negligible, *b* = -0.03, 95% CI [-0.09, 0.03]. On the other hand, there was a direct effect of display type on location accuracy, with the semantic display type predictive of poorer location accuracy, *b* = -0.18, 95% CI [-0.36, -0.00]. A parallel exploratory mediation analysis using Euclidean distance as the outcome variable found similar results: the semantic display type was predictive of shorter study times, *b* = -0.07, 95% CI [-0.14, -0.00], and longer study times were associated with shorter distances to the target, *b* = -0.13, 95% CI [-0.21, -0.05]. The effect of display type on Euclidean distance was, again, not mediated by study time, *b* = 0.01, 95% CI [-0.02, 0.04], while there was a direct effect of display type on Euclidean distance, with the semantic display type leading to participants choosing further distances from the target, *b* = 0.10, 95% CI [0.00, 0.19]. Figure 6 shows the estimated regression coefficients for the mediation models.

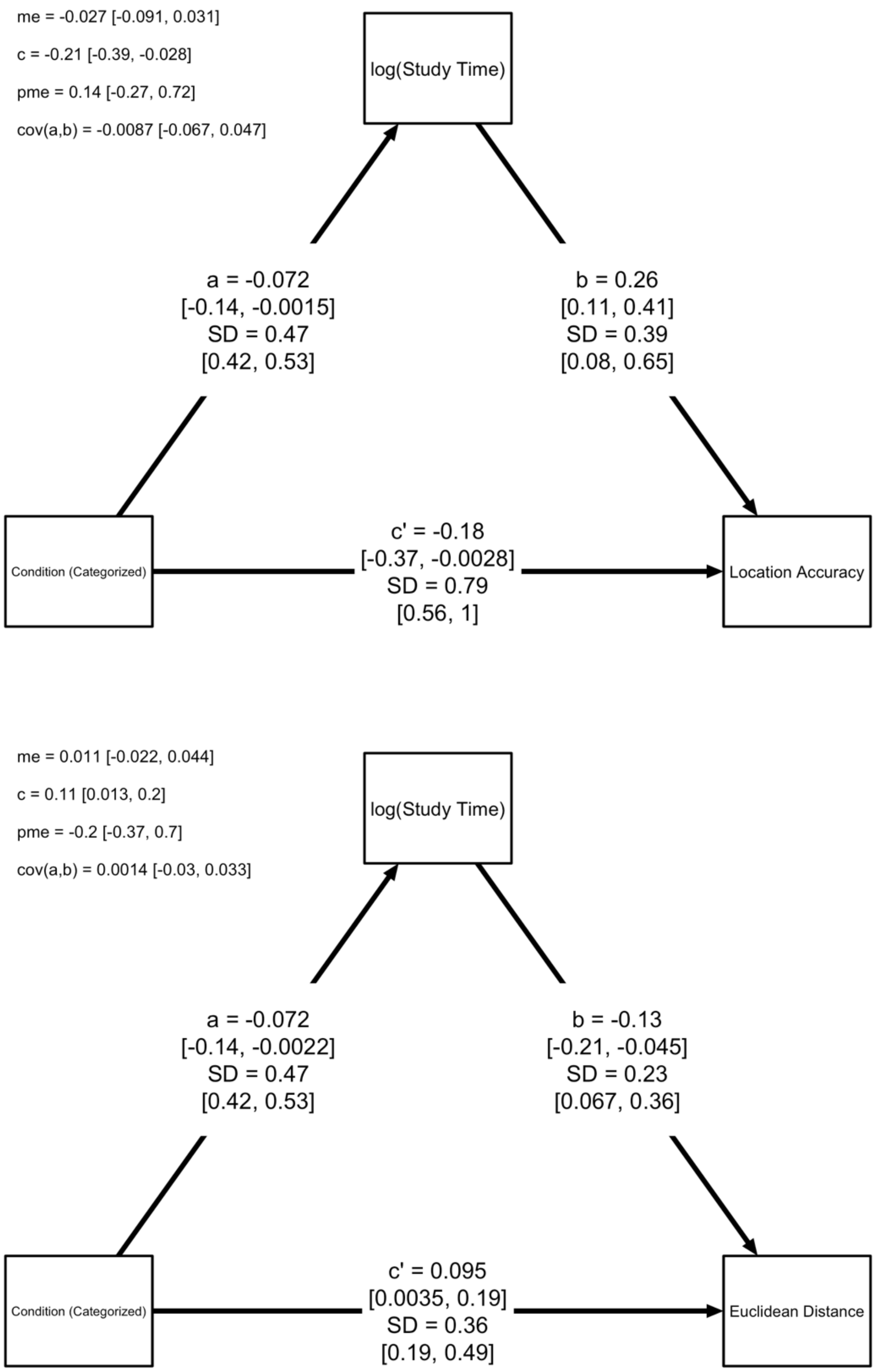


Figure 6. Estimated standardized regression coefficients for the relationship between condition and location accuracy as mediated by study time (top) and the relationship between condition and Euclidean distance as mediated by study time (bottom). me = mediated effect, c = total effect, c' = direct effect, pme = proportion mediated effect.

Item Location Substitution Rate

An exploratory mixed-effects logistic regression showed that participants were not more likely to substitute another item location when making an error in the semantic compared to scrambled grid, *b* = 0.02, 95% CI [-0.06, 0.10], *z* = 0.47, *p* = .635.

Discussion

In Experiment 3, I replicated the finding that participants performed worse in a location memory task when presented with a display consisting of items from a single category compared to a scrambled display consisting of items from different categories. Specifically, I found that overall location accuracy significantly decreased in the semantic display, and the average distance between the chosen location and the target location was greater in the semantic display. While participants allocated more time to studying the items from different categories compared to the items from a single category, this did not account for the cost of relatedness on location memory performance. The effect of display type (item relatedness) on location accuracy was not mediated by study time, and nor was its effect on Euclidean distance. Together with the earlier experiments, these results provide evidence that semantically related items lead to a decrease in participants' ability to correctly recall item locations. However, I did not find that participants were more likely to substitute another item location when making an error in the semantic compared to scrambled grid, suggesting that the semantic display did not increase the likelihood that participants confused whole item locations with each other.

General Discussion

Across four experiments, I demonstrated that participants tended to perform worse in a location memory task when presented with a display consisting of items from a single category compared to a scrambled display consisting of items from different categories. In Experiment 1a, while semantic relatedness increased item recall performance, it decreased location memory accuracy; in Experiment 1b, the latter result was numerically in the same direction but was not significant. In Experiment 2, using a stronger condition manipulation and a different set of item categories, I replicated the finding that semantic relatedness significantly decreased location memory accuracy. In Experiment 3, I replicated the cost of semantic relatedness in location memory accuracy as well as in Euclidean distance with a self-paced study task and a different set of item categories. While I found that semantic relatedness reduced study time, I did not find that cost of semantic relatedness on location memory was mediated by this reduced study time. Taken together, these results demonstrate a robust cost of semantic relatedness to item location memory, in contrast to its benefit to item recall memory. Thus, the current results are an example of the "double-edged" sword of semantic relatedness in memory: depending on the situation, semantic relatedness can lead to benefits and/or costs, both within and across different tasks (Kahana et al., 2022; Nelson et al., 2013).

I have considered two possible explanations for the cost of semantic relatedness to location memory. According to the semantic interference hypothesis, while increased semantic similarity across items (generalization) increases the likelihood of outputting an item in free recall, it also increases interference across these items, reducing performance when discrimination among the to-be-remembered items is required. Unlike a free recall task, correct performance in a location memory task is contingent upon successful discrimination between items. A strong version of this argument is that semantic similarity-based interference increases the likelihood that another item is erroneously recovered and recalled in the position of the targeted one (Tse, 2009). I investigated the confusion-based hypothesis by examining the item location substitution rate (the likelihood of substituting a different item's location for the target location) for the semantic and scrambled displays. However, across four experiments, I found little evidence that participants were more likely to substitute another item location when making an error in the semantic compared to scrambled grid (nonsignificant Experiment 1a, 1b and 3, and a small effect in Experiment 2). Overall, these results would suggest that the semantic display did not appear to increase the likelihood that whole item locations would be confused with each other. However, the results do not entirely preclude a role for interference: for example, increased semantic-based competition at output might reduce the likelihood that any one item's location is recalled, which could increase location errors but not item substitution errors specifically. Future research should further examine the role of semantic similarity-based interference and its potential role in the cost to location memory.

A second hypothesis I considered was a metacognitive loafing hypothesis: that the cost might be driven by participants spending less time and/or effort studying the items in the semantic display compared to the scrambled display. In Experiment 3, I investigated this possibility by allowing participants to self-pace their study time during the location memory task. While I did find that the semantic display was associated with decreased study time, the semantic cost to location memory was not mediated by study time, and I continued to observe a direct effect of display type on location memory. These results cannot entirely rule out the possibility that metacognitive loafing might play a role in the cost of semantic relatedness to location memory: self-paced study time may not fully capture participants' experiences of subjective effort. However, to the extent that participants did spend less time studying the items in the semantic display, I did not find that the reduced study time was able to explain the cost to location memory observed in that condition. Thus, across four experiments I have established the existence of a cost to location memory in semantically homogenous displays, as well as tested two potential mechanisms, seemingly ruling out strong versions of both as compelling explanations for that cost. I anticipate that future explorations into the mechanisms behind the cost will prove to be productive avenues for further research.

Another contribution of the current work is with respect to participants' subjective preferences and ratings. In Experiments 1a and 1b, participants rated the semantic display as more difficult than the scrambled display. These judgments likely reflected the objective location memory performance differences between the two conditions. While semantically related word lists are usually judged to be easier to learn than unrelated word lists in standard list learning tasks (Hourihan & Tullis, 2015; Matvey et al., 2006), objective performance also tends to be better with the related lists, so participants' subjective judgments do not conflict with objective performance. In the current experiments, however, the semantically related set of words was associated with poorer performance. Participants were indeed sensitive to their performance to some degree: in Experiment 1a/1b, they rated the semantic display as more difficult than the scrambled display, and their subjective ratings of difficulty were predicted by their objective performance. However, participants also judged the semantic display to be the easier display in higher proportions than would be expected from their objective performance. In Experiment 2, after controlling for objective performance, I found that participants tended to prefer the semantic display to the scrambled display when explicitly asked to choose one of the two display types for a future task. Thus, participants appear to have a metacognitive bias towards the semantic display—that is, they perceive the semantically related item display to be easier (Experiment 1a/b) or indicate it as preferred (Experiment 2) more than they objectively should, based on performance considerations. These results suggest that people may have some sort of a pre-existing notion that semantically related items are easier to learn in general, though experience with a large semantic cost was able to shift participants' preferences to some degree. Consistent with the idea that participants exhibit a semantic preference/bias, in Experiment 3, I found that participants spent less time studying the item locations in the semantic display compared to the scrambled display, suggesting that they may anticipate these items to be easier to learn. This is remarkable given that I avoided soliciting participants' preferences in this experiment to avoid leading/biasing them in one direction or another. I anticipate that relatedness effects on location memory will prove a fruitful avenue to examine potential metacognitive biases.

Finally, what implications might the current results have for location memory in the wild? In real-life environments, we often place items according to functional and/or semantic considerations (e.g., tools are all kept together in the shed, while clothes are kept in the closet). While this type of placement might benefit our ability to recall the items themselves, our ability to recall specific item-location associations appears to be rendered worse. While the current studies used carefully controlled word stimuli in a sparse grid, future investigations may investigate whether the results generalize across more ecologically valid contexts (e.g., using real objects in physical space, or virtual items in virtual reality space).

Conclusion

While semantic relatedness tends to benefit memory, it may also lead to memory costs in certain contexts. Here, I have shown that the semantic relatedness of words within a display had the effect of increasing item recall performance but decreasing location memory accuracy. These results are consistent with semantic relatedness serving as a “double-edged sword” in memory. Relatedness improves memory recall by increasing the likelihood of target retrieval, but worsens location memory performance which requires distinguishing between similar target representations.

Combined analysis: Relatedness cost on objective location memory

Chapter 1 ended on a somewhat inconclusive note for the distinctiveness/interference hypothesis, as the effect of relatedness on one direct measure of increased inter-item interference (i.e., the item substitution rate) was not found to be robust. Here, I present the results of an analysis on the combined data from Chapter 1 (*N* = 527) and Chapter 3 (*N* = 828), which gives an overall estimate of the relatedness effect on location memory.

Table 1. Mean (SD) objective memory measures on Chapter 1 and 3 combined data (*N* = 1355)

	Related	**Unrelated**
Location Accuracy	.42 (.24)	.45 (.25)
Euclidean Distance	1.33 (0.69)	1.23 (0.69)
Item Substitution Rate	.44 (.24)	.43 (.25)

Location Accuracy. A mixed-effects logistic regression revealed a significant main effect of relatedness, such that location memory accuracy was lower for the related items than unrelated, *b* = -0.12, 95% CI [-0.19, -0.06], *z* = 3.58, *p* < .001. Participants were less likely to select the correct item location when the items were related.

Euclidean Distance. A mixed-effects linear regression revealed a significant main effect of relatedness, such that the mean Euclidean distance between selected and target locations was greater for the related items than unrelated, *b* = 0.08, 95% CI [0.04, 0.12], *t* = 4.12, *p* < .001. Participants' answers tended to be locations that were further away from the target location when the items were related.

Item Substitution Rate. A mixed-effects logistic regression revealed a significant main effect of relatedness: participants were more likely to select another item on incorrect trials for the related items display, *b* = 0.07, 95% CI [0.01, 0.13], *z* = 2.09, *p* = .036. Participants were more likely to substitute a related item location when making an error, suggesting that the related items were more easily confused with each other (increased inter-item interference) compared to the unrelated items.

The combined analysis shows a robust memory cost of relatedness: participants performed worse in the location memory task when the items were related, whether using a binary measure of accuracy or a continuous measure of distance. In support of the interference hypothesis, the item substitution rate was also found to be significant. This effect was not large, however, suggesting that only a fraction of the cost was driven by whole-item level confusions.

Chapter 2

People usually predict that they will remember related words better than unrelated words. This relatedness benefit in metamemory judgments is thought to be mediated in part by participants' *a priori* beliefs that related words are easier to remember. This benefit tends to be metacognitively accurate in the sense that relatedness does benefit memory in most cases. Here I investigated the effect of relatedness on participants' memory predictions in a location memory task, where the locations of related words are known to be remembered worse than those of unrelated words. Participants were presented with a vignette that described a location memory task where words would be presented in random locations within grid displays. In one condition, the grid display contained words that were all from the same category (related); in the other condition, the grid display contained words that were all from different assorted categories (unrelated). Across six experiments, participants consistently predicted that they would remember more correct locations when the words were related than when they were unrelated – a "relatedness halo". I advance an explanation of the results based on the idea that the experimental context (i.e., the task vignette and prompt) activates different naïve theories about memory, with mechanisms that lead to a benefit being more readily cued and activated. These results suggest that metamemory beliefs are constructed online from available and accessible information that varies depending on the context of its elicitation.

Metamemory refers to our knowledge and beliefs about memory, and the ability to evaluate and control our own memory performance. Metamemory processes play an important role in everyday decision making, such as in deciding whether to continue studying for a test, or whether one should write down a reminder for a future appointment. Theoretical accounts of metacognition, such as the dual-basis view (Koriat, 1997; Koriat et al., 2004) and analytic processing theory (Mueller & Dunlosky, 2017), emphasize a central role for *a priori* theories or beliefs about memory in our metacognitive judgments. Yet the precise nature of these beliefs remains relatively underexplored. As previous research has shown, participants' expressed beliefs about memory can be dissociated from their item-by-item judgments of learning (e.g., Koriat, et al., 2004; Kornell et al., 2011). It is not clear when and why participants rely on beliefs in their judgments, suggesting the need for a coherent theory of metamemory beliefs (Koriat, et al., 2004). In the present investigation, I further examine the nature of metamemory beliefs about semantic relatedness.

Semantic relatedness is an important factor that influences metamemory judgments (e.g., Arbuckle & Cuddy, 1969; Castel et al., 2007; Koriat, 1997; Koriat et al., 2004; Lu et al., 2022; Mueller et al., 2013). Investigations into the effects of relatedness on predictions of learning have relied heavily on the paired associates learning task (Arbuckle & Cuddy, 1969), wherein participants tend to give higher judgments of learning to related word pairs (e.g., *cow-milk*) than to unrelated word pairs (e.g, *fish-pen*), although similar results have been obtained in predictions of free recall for related word lists compared to unrelated word lists (e.g., Lu et al., 2022). Researchers have generally considered two explanations for this relatedness effect. The first asserts that the relatedness effect reflects participants applying an *a priori* belief about relatedness being beneficial for memory (e.g., Mueller et al., 2013; Soderstrom & McCabe, 2011); the second asserts that the relatedness effect reflects the more fluent experience of processing the related word pairs compared to the unrelated word pairs (i.e., increased processing fluency; Castel et al., 2007; Undorf & Erdfelder, 2015). Evidence that the relatedness effect is, at least in part, belief-based comes from investigations that have attempted to eliminate the influence of processing fluency. For example, Mueller et al. (2013) had participants read a description about a hypothetical paired associates learning task involving related and unrelated word pairs. Participants then estimated how many related and unrelated pairs would be recalled. Their estimations were higher for related word pairs than for unrelated word pairs (68% vs. 36%), indicating that they had an *a priori* belief about the relative memorability of related and unrelated word pairs.

The results of Mueller et al. (2013) could be taken to reflect either a general belief about the beneficial effect of relatedness on memory performance (e.g., "relatedness benefits memory in general") or a more task-specific belief (i.e., "relatedness specifically benefits performance in this memory task"). Given that beliefs about semantic relatedness have to date been investigated in tasks where there is a benefit (e.g., paired associate learning; free recall), it is impossible to distinguish between these possibilities. Indeed, adjudicating between them is especially difficult precisely because semantic relatedness is beneficial to memory in many cases. Recently, however, I have found that relatedness is associated with a cost when individuals are tasked with remembering the locations of items (Lu et al., 2023). In this location memory task, participants are presented with to be remembered items on a grid display. In one condition, the grid display contained words that were all from the same category (related items); in the other condition, the grid display contained words that were all from different assorted categories (unrelated items). Across multiple experiments, Lu et al. (2023) reported that memory for the item locations was consistently worse in the related-items condition than in the unrelated-items condition. Critically, only memory for locations was rendered worse; consistent with previous research, relatedness had a positive effect on item memory as measured by a free recall task (Lu et al., 2023). The location memory task thus provides a unique opportunity to investigate beliefs about relatedness, both in a context where it has an objective benefit for memory (item memory) and in a context where it has an objective cost to memory (location memory). In the current investigation, I solicited participants' beliefs about how item relatedness would influence memory in this task (Koriat et al., 2004; Kornell et al., 2011; Mueller et al., 2013).

Overview of Current Investigation

I report six pre-registered experiments (Experiment 4: https://osf.io/e86f5; Experiment 5a: https://osf.io/39kfn; Experiment 5b: https://osf.io/25jfa; Experiment 6: https://osf.io/shm4x; Experiment 7: https://osf.io/89neq; Experiment 8: https://osf.io/t85gq) that solicited participants' predictions of how item relatedness would influence their performance in a location memory task. In each experiment, participants were given a vignette that described a hypothetical memory experiment where words to be remembered would be presented within grid displays (e.g., Lu et al., forthcoming; Lu et al., 2023). In one condition, the grid display would contain words that were all from the same category (Related); in the other condition, the grid display would contain words that were all from different assorted categories (Unrelated).

In Experiment 4, participants were either told that the memory test would be a location memory test (Location Memory group) or an item recall test (Item Memory group). I anticipated that participants would predict a benefit of relatedness for item memory, as has been previously observed for free recall (e.g., Lu et al., 2022; 2023). On the other hand, no previous studies have investigated whether this belief would continue to apply in a context like location memory, where relatedness leads to a memory cost. As noted, one possibility is that participants may apply a general belief that relatedness is always beneficial to memory, in which case we should expect them to predict a relatedness benefit of a similar magnitude in both the location memory task and the item memory task. A number of studies have indeed suggested that individuals might have naïve beliefs about the influence of relatedness on memory (e.g., Carroll et al., 1997; Koriat & Bjork, 2005; 2006).

In addition, in my previous investigations (Lu et al., 2023), there was some evidence that individuals might indeed think that relatedness was beneficial even when remembering locations. For example, when asked to indicate which of the two displays was easier to learn in a post-task questionnaire (Lu et al., 2023, Experiment 1a/b), participants were rather evenly split between related and unrelated item displays, despite most of them performing worse in the former condition. Furthermore, during a self-paced version of the location memory task (Lu et al., 2023, Experiment 3), participants spent less time studying the related items compared to the unrelated items, which again might reflect a general "relatedness is beneficial to memory" belief. However, since all of this previous research solicited judgments only after encoding the items, it could also reflect the contributions of processing fluency: related items might feel easier to encode (even if there is a cost at retrieval). A second possibility is that participants hold more nuanced, task-specific beliefs about the influence of relatedness on memory such that their predictions for the location memory and item memory tasks might diverge (e.g., by predicting a cost of relatedness for location memory).

Experiment 4

Method

Participants

Data from 400 participants (256 women, 138 men, 5 other, 1 unknown, *M* = 39.30 years, *SD* = 12.93) were analyzed. A power calculation performed with Superpower's *ANOVA_exact* (Lakens & Caldwell, 2021) estimated that this sample size would be sufficient to achieve 90% power for the relatedness effect in the location group[2]. Participants were recruited from Prolific and compensated GB£0.25 for approximately two minutes.

Materials and Procedure

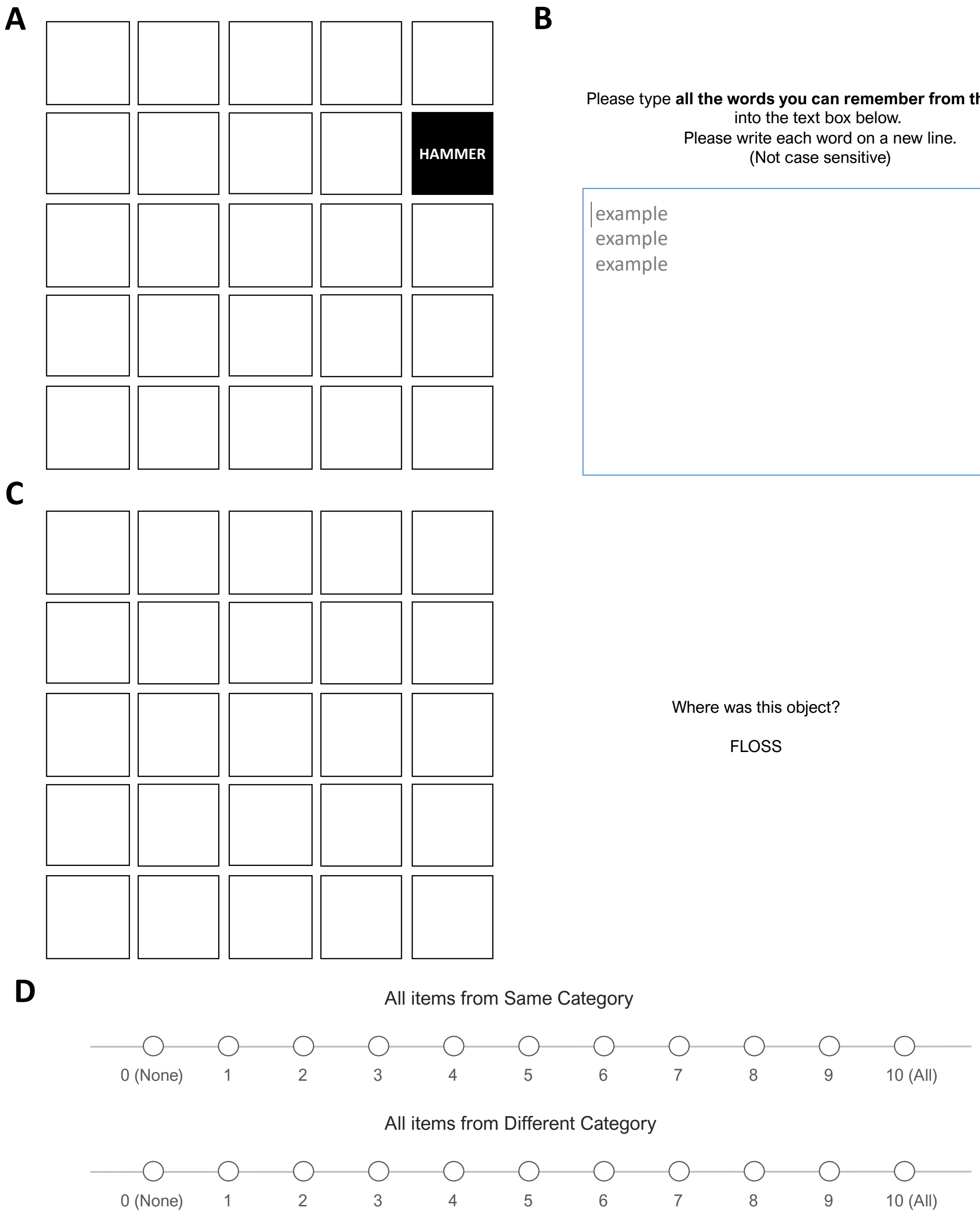
Participants were randomly assigned to either the item memory group (*N* = 200) or the location memory group (*N* = 200). After being given a vignette describing the experimental task they were to perform, participants were asked to make a prediction of their performance, and then asked to report any reasons for their predictions.

Experimental Vignette. The experimental vignettes were adapted from Lu et al. (2023). Participants were given the following description, and some example pictures of the encoding and test tasks (see Figure 7A-7C):

> "In this task, you will be presented with 10 words in a grid display (see below). In the learning phase, each word will appear on the grid one at a time. Please try to remember [each word/the location of each word] that was presented.
>
> Once you have seen all of the words, your memory will be tested. In the memory test, you will be given [a blank text box and asked to recall all the words that were presented

[2] The following estimates were used: M_Item_Related = 5.90, M_Item_Unrelated = 4.54, M_Location_Related 4.88, M_Location_Unrelated = 4.49, within-subjects *r* = 0.46, *SD* = 1.62. These were obtained from previous experiments conducted in the lab.

to you/ each target word one at a time and you will have to indicate its location in the grid]."



Please type **all the words you can remember from the display** into the text box below.
Please write each word on a new line.
(Not case sensitive)

Where was this object?

FLOSS

D
All items from Same Category
0 (None)
1
2
3
4
5
6
7
8
9
10 (All)
All items from Different Category
0 (None)
1
2
3
4
5
6
7
8
9
10 (All)

Figure 7. (A) An illustration of the encoding phase provided to participants. On screen, this was presented as an animated gif showing that different items would appear one by one within the display. (B) An illustration of the item memory test and (C) of the location memory test. (D) Rating scales for related vs. unrelated conditions.

Related and Unrelated Memory Prediction. After reading the vignette, participants were then given the following instruction and asked to predict their performance on each type of display (Related and Unrelated) on a scale ranging from 0 (None) to 10 (All):

> "**Please read these instructions carefully.**
>
> As noted previously, you will complete a task that involves learning [words/the locations of words] presented in a grid display. Some displays contain words that are all in the same category (e.g., all words are kinds of FURNITURE) and some displays contain words from different assorted categories (e.g., one word might be a kind of FURNITURE, one word might be a TOOL, another word might be a TOY).
>
> For each type of display, on the scales below please estimate the number of words (out of 10) [which you will be able to correctly recall/for which you will be able to correctly recall the location]."

Self-Reported Reasoning. After giving their predictions, on the next screen participants were reminded of them and asked to type their reasoning into an on-screen text field:

> "In the last task we asked you to make predictions about the number of words you would be able to [correctly recall/correctly recall the location]. You estimated [X] when all items are from the same category, and you estimated [Y] when all items are from different categories. Please tell us about your reason(s) for selecting the number of items for each display type."

Finally, participants were asked to provide their age and gender and to complete an attention and effort check questionnaire. They were debriefed that they would not be required to actually complete the memory task, and additional consent to use their data was sought following this debriefing.

Results

Data from 24 participants were not analyzed according to the exclusion criteria set in the pre-registration, leaving $N = 400$ after replacement. ANOVAs were performed using the *afex* package in R (Singmann et al., 2015); ANOVA effect sizes reported are generalized eta squared. Data and analysis code for all experiments are available at https://osf.io/bvyef/files/.

Effects of relatedness and task on memory predictions

A 2 (Relatedness: Related vs. Unrelated Items) x 2 (Memory Type: Item vs. Location) mixed ANOVA revealed a main effect of Relatedness, $F(1,398) = 53.96$, $MSE = 2.00$, $p < .001$, $\eta_G^2 = .038$, with memory predictions higher for related items than for unrelated items. There was also a main effect of Memory Type, $F(1,398) = 18.79$, $MSE = 4.92$, $p < .001$, $\eta_G^2 = .032$: Item memory predictions were higher than location memory predictions. The interaction was significant, $F(1,398) = 6.00$, $MSE = 2.00$, $p = .015$, $\eta_G^2 = .004$. For item memory predictions, related items were predicted to be remembered better than unrelated items, $F(1,199) = 50.04$, $MSE = 1.92$, $p < .001$, $\eta_G^2 = .071$. For location memory predictions, related item locations were also predicted to be remembered better than unrelated item locations, $F(1,199) = 11.51$, $MSE = 2.09$, $p < .001$, $\eta_G^2 = .016$, although this effect was significantly smaller than that for item memory predictions. Table 2 shows the mean memory predictions in each condition.

Table 2. Experiment 4: Mean memory predictions (SD) by item relatedness and memory type

	Related	**Unrelated**
Item Memory (*N*=200)	5.83 (1.79)	4.85 (1.77)
Location Memory (*N*=200)	4.90 (2.05)	4.41 (1.83)

Effects of task on relatedness belief direction

The above analyses showed that participants' aggregate predicted relatedness benefit was greater for the item task than for the location task. However, the aggregate results could mask important differences in participants' beliefs as well as in the distribution of those beliefs.

Based on their related vs. unrelated prediction differences, I conducted exploratory analyses after sorting participants into three groups: (1) those who predicted that they would remember more items/item locations for the related items (related benefit), (2) those who predicted that they would remember fewer items/item locations for the related items (related cost), and (3) those who predicted that they would remember equal numbers for related vs. unrelated items/item locations (no difference). A chi-square test suggested that the proportions in each group differed for location vs. item memory, $\chi^2(2) = 21.06$, $p < .001$. Pairwise comparisons with Bonferroni-adjusted *p*-values revealed that there were significantly more participants predicting a benefit for item memory compared to location memory, $\chi^2(1) = 9.62$, $p = .004$, and significantly more participants predicting a cost for location memory compared to item memory, $\chi^2(1) = 16.66$, $p < .001$, whereas the proportion of participants predicting no difference did not differ, $\chi^2(1) = 0.09$, $p = .758$. Interestingly, for the subset of participants who predicted a relatedness benefit, the size of the predicted benefit did not differ for item memory ($M = 2.04$, $SD = 1.59$, $N = 113$) vs. location memory ($M = 2.19$, $SD = 1.48$, $N = 81$), $F(1,192) = 0.39$, $MSE = 2.39$, $p = .532$, $\eta_G^2 = .002$. The size of the predicted cost also did not differ for item memory (*M*

= 3.18, *SD* = 2.96, *N* = 11) vs. location memory (*M* = 2.03, *SD* = 2.03, *N* = 39), *F*(1,48) = 2.25, *MSE* = 5.10, *p* = .140, η_G^2 = .045, although the small *Ns* warrant caution.

Although the ANOVA results left open the possibility that individuals hold the same belief that relatedness is beneficial (just to a lesser extent) in the location task, this analysis suggests that the interaction is driven by more participants predicting a benefit for item, and more participants predicting a cost for location (i.e., a difference in belief kind rather than degree). For participants anticipating an effect of relatedness for either task, the benefit or cost was comparable in magnitude. Table 3 shows the proportion of participants in each group.

Table 3. Experiment 4: Proportions of participants who predicted a related benefit, related cost, or no difference for item vs. location memory

	Related Benefit	**Related Cost**	**No Difference**
Item Memory **(*N*=200)**	56.5%	5.5%	38.0%
Location Memory **(*N*=200)**	40.5%	19.5%	40.0%

Effects of task and relatedness on self-reported prediction reasoning

Participants' self-reported reasons for their predictions were classified as belonging to various categories as outlined in Table 4. The coding scheme was developed iteratively, initially based on *a priori* theory (Lu et al., 2023; Lu et al., 2024) and further modified after reading through participants' responses. There were two independent coders who were blind to participants' predictions and condition assignment, and each reason could be assigned to multiple categories under the coding scheme. As a measure of inter-rater reliability, I calculated

Cohen’s kappa (κ) for each category, a measure which corrects for chance agreement between raters (Cohen, 1960). Cohen’s κ ranged from 0.61 to 0.84, suggesting substantial to almost perfect agreement between the two raters (Cohen, 1960).

Table 4. Experiments 4, 7 and 8: Coding scheme for participant’s reasoning

Belief Classification	Description	Example
Mental Imagery, Visualization	Participant indicates that they can more easily form mental imagery or visualize the items for one of the displays	*It will be easier to recall items from the same category as they can be visualized in one room*
Associations, Connections, Cueing	Participant indicates that they can more easily form links between items for one of the displays	*I imagine that it will be easier to link words from the same category in my mind*
Distinctiveness, Interference	Participant indicates that they can more easily distinguish between the items for one of the displays	*Items from the same category are less likely to be unique and memorable compared to different categories which stand out*
Unspecific Related Preference	Participant reiterates that they find it easier to remember related words, but no particular reason provided	*It is easier for me to remember words when they are related to each other*

Unspecific Unrelated Preference	Participant reiterates that they find it easier to remember unrelated words, but no particular reason provided	*The items from different categories may be easier to remember*
Indifference	Participant indicates that the conditions will have little influence on memory	*I'm not sure the same or different category will matter*
Memory Ability	Participant indicates that their performance is dependent on their memory ability	*My memory is usually quite good*
Other Reason	Participant provides an unlisted reason for preferring either of the displays	*I will try to make a story with the related words*

Table 5 shows the percentages of participants expressing a particular type of reason (as coded by the primary coder). Of note, participants predicting a benefit tended to reason that it would be easier to make associations between related items, or else they simply reiterated this belief without additional explanation or support, for either the item or the location memory task. In contrast, participants predicting a cost of relatedness in the location memory task often reasoned that the distinctiveness of the different-category items would be beneficial. Finally, participants predicting no effect of relatedness tended to reiterate their belief that relatedness would not matter and/or referred to their memory ability as the determining factor.

Table 5. Experiment 4: Percentages of participants' self-reported reasoning behind their predictions. Note that responses could be classified into multiple reasons, and some were unclassifiable/unusable, hence columns do not sum to 100%.

	Item Memory (*N*=200)			**Location Memory (*N*=200)**		
	Related Benefit	Related Cost	No Difference	Related Benefit	Related Cost	No Difference
Imagery	1.8%	-	-	7.4%	5.1%	-
Association	33.6%	-	-	24.7%	2.6%	1.3%
Distinctiveness	0.9%	9.1%	-	2.5%	53.8%	-
Unexplained Related Preference	34.5%	18.2%	1.3%	46.9%	5.1%	1.3%
Unexplained Unrelated Preference	-	9.1%	-	-	15.4%	-
Indifference	1.8%	9.1%	42.1%	6.2%	7.7%	52.5%
Memory Ability	32.7%	27.3%	64.5%	24.7%	35.9%	42.5%
Other Reason	12.4%	-	1.3%	6.2%	2.6%	-

Discussion

Overall, participants predicted that related items would be better remembered than unrelated items, both for the item memory test and for the location memory test. However, this effect was clearly modulated by task, such that the relatedness benefit was predicted to be smaller overall in the location memory task. By sorting participants into those that believed in a relatedness benefit, cost, or no difference, I found that the task by relatedness interaction appeared to be driven by relative differences in the direction of beliefs. While most individuals in both tasks thought that relatedness would benefit memory, there were more such individuals in the item memory task, whereas more individuals in the location memory task predicted a cost (and about an equal proportion of individuals in both tasks predicted no difference). Among people who believed in a relatedness benefit or cost for either condition, the size of the estimated effect was similar for item memory and location memory.

These results paint a nuanced picture. First, there appears to be at least some degree of belief in a general benefit of relatedness for memory: Overall, participants predicted a relatedness benefit in both item memory and location memory tasks. In addition, for participants that expressed a belief in a relatedness benefit, the magnitude of that benefit was comparable across tasks. This seems consistent with the idea that a large subset of participants believe in a kind of general relatedness benefit that is not moderated by task (at least by the tasks used here), and seems inconsistent with an alternative explanation wherein individuals believe that relatedness is beneficial but just less so in the location memory task. The overall reduction in the influence of relatedness for location memory predictions appears to be driven by a sizable minority of participants who predicted a cost of relatedness in the location memory task. That is, there appear to be individual differences in people's metamemory beliefs about relatedness: Some participants applied the belief that "relatedness benefits memory" to the location memory task whereas others concluded that relatedness would result in a cost in that task. Participants' self-reported reasoning was also illuminating: People predicting a relatedness benefit often invoked the idea of making more connections/associations between the related items, whereas a majority of people predicting a relatedness cost in the location memory task attributed this to the notion of reduced distinctiveness in the related condition (or increased distinctiveness in the unrelated condition).

The current results suggest that participants hold at least two distinct *naïve theories* or *mental models* about how relatedness influences memory. According to one, relatedness benefits memory by creating associations between items, making each item easier to retrieve. According to another, relatedness impairs memory by making individual items less distinct. The first naïve theory (resulting in *relatedness-as-benefit* predictions) appears to be much more prevalent, at least in the current context: In the aggregate, there was a benefit predicted in both tasks, and a sizeable proportion of participants predicted that there would be a benefit. Since more participants predicted a relatedness benefit for the item task, this suggests that the item task description is more likely to lead to retrieval of the naïve relatedness-as-benefit theory and its associated mechanisms. More participants predicted a relatedness cost for the location task than for the item task, and although these participants were still a minority, this suggests that the location task description was comparatively more likely to lead to retrieval of naïve theories pertaining to the cost.

The results also suggest, however, that once a given theory is retrieved and activated, the magnitude of the predicted benefit is comparable across tasks. Thus, one way to explain the pattern of results that I observed is to suggest that the features of the vignette activate an internal representation of the naïve theory that determines the direction of the impact. Participants are more likely retrieve a relatedness-as-cost mechanism when considering the demands of the location memory test (that it would require distinguishing between different items). Once a particular theory or mechanism is retrieved, however, it is translated onto the scales in a manner that is insensitive to how it was initially activated. Finally, a substantial proportion of participants predicted that relatedness would not influence memory, and their numbers were consistent across the two tasks. This suggests a potential third distinct naïve theory centering around memory malleability and flexibility, with an individual's perceived memory ability as the limiting factor. These participants believe that their memory performance is what is critical and that it remains relatively unchanged in the face of experimentally manipulated factors such as relatedness.

While I had anticipated that participants would predict that relatedness would benefit item memory, the finding of a predicted benefit in the location memory group—against the background of a known memory cost (Lu et al., 2023)—was more unexpected. In the next two experiments, I replicate the finding of a relatedness benefit in the location memory condition and rule out some alternative explanations.

Experiments 5a and 5b

An alternative possibility that I considered for the overall location memory benefit in metamemory was that a significant subset of participants in Experiment 4 may have misunderstood the location memory test to also involve recalling the items.[3] While I thought this unlikely, given that all participants were given a detailed description of the grid display and memory task, subsequent experiments included an additional reminder that the test would be about item locations and that the items would be provided at test.

Also, in Experiment 4, the related items condition was always described first in the vignette, and the related/unrelated Likert rating scales on the screen always had the related item

[3] The reverse may be more likely: Given the provided description and picture of the items being presented in grid displays, participants in the item memory condition may have thought that their task was to remember the locations. These "task confusions" may have resulted in an underestimation of the benefit of relatedness for item memory.

judgment above the unrelated item judgment (see Figure 7D), both of which could have indirectly biased participants to give higher ratings to the related condition. Therefore, in Experiment 5a, the orders of the related and unrelated condition descriptions as well as the presentation order of the related/unrelated Likert rating scales on screen were manipulated between participants. Experiment 5b was an identical replication of 5a, focusing exclusively on the condition that was thought most likely to bias participants toward the unrelated condition (i.e., unrelated condition description first and unrelated scale first).

Method

Participants

In Experiment 5a, participants again were recruited from Prolific and compensated GB£0.25 for two minutes. After exclusions, I analyzed complete data from 400 participants (215 women, 176 men, 5 other, 4 unknown, *M* = 37.84 years, *SD* = 13.31). In Experiment 5b, participants were University of Waterloo undergraduates who were compensated with course credit. I pre-registered an end-of-school term stopping rule, allowing us three weeks to collect data, with a minimum goal of *N* = 200. The final sample size was *N* = 471 without exclusions (311 women, 150 men, 3 other, 7 unknown, *M* = 19.96 years, *SD* = 3.39) and *N* = 336 after exclusions (221 women, 113 men, 1 other, 1 unknown, *M* = 20.00 years, *SD* = 3.64). A large number of participants had to be excluded because they self-reported poor attention during the task. This was surprising given that the study duration was very short, so I analyzed the data both before and after exclusions and confirmed that they were highly similar. Below I report the set of analyses after the pre-registered exclusions (*N* = 336); the analyses before exclusions can be found on https://osf.io/bvyef/files/.

Materials and Procedure

All experiments followed a procedure almost identical to the location memory condition in Experiment 4 except that participants were given an additional reminder (that the test would be on item locations) within the pre-task prediction instructions as follows:

> "**Please read these instructions carefully.**
>
> As noted previously, you will complete a task that involves learning the **locations of words** presented in a grid display. In the memory test, you will be given each word one at a time and you will have to indicate its location in the grid."

In Experiment 5a, the order of related/unrelated condition descriptions in the vignette was manipulated between-participants. This resulted in four possible order combinations: related description first, related scale first (R-R); related description first, unrelated scale first (R-U); unrelated description first, related scale first (U-R); unrelated description first, unrelated scale first (U-U). In Experiment 5b, all participants were assigned to the unrelated description first, unrelated scale first (U-U) order. The final self-reported reasons probe was omitted in these three experiments.

Results

In Experiment 5a, following the pre-registered exclusion criteria, data from 28 participants were not analyzed, leaving *N* = 400 after replacements. In Experiment 5b, following the pre-registered exclusion criteria, data from 135 participants were not analyzed, leaving *N* = 336 at the conclusion of the stopping rule.

Effects of relatedness on memory predictions

In Experiment 5a, a 2 (Relatedness: Related vs. Unrelated Items) x 2 (Description Order: Related First vs. Unrelated First) x 2 (Scale Order: Related First vs. Unrelated First) mixed

ANOVA revealed no main effects of description or scale order, and no interactions of description or scale order with relatedness (all *p*s > .646). There was, however, a main effect of relatedness, such that participants predicted that the locations of related items would be remembered better than those of unrelated items, *F*(1,396) = 6.37, *MSE* = 2.83, *p* = .012, η_G^2 = .005.

I replicated this result in Experiment 5b: A paired-samples t-test revealed that participants predicted that the locations of related items (*M* = 5.68, *SD* = 2.05) would be remembered better than those of unrelated items (*M* = 4.76, *SD* = 1.95), *t*(335) = 8.32, *p* < .001, *dz* = 0.46 [0.31, 0.62], in the condition seemingly most likely to bias individuals away from such a belief. Table 6 shows the mean location memory predictions in each condition in Experiment 5a.

Table 6. Experiment 5a: Mean location memory predictions (SD) by item relatedness, description, and scale order

	Description Order	**Scale Order**	**Related**	**Unrelated**
R-R (*N*=100)	Related First	Related First	4.96 (2.01)	4.51 (1.88)
R-U (*N*=101)	Related First	Unrelated First	4.88 (1.99)	4.69 (1.95)
U-R (*N*=99)	Unrelated First	Related First	4.94 (2.11)	4.64 (2.06)
U-U (*N*=100)	Unrelated First	Unrelated First	4.99 (2.23)	4.73 (2.14)

Relatedness belief direction

As in Experiment 4, I sorted participants into those that believed in a relatedness benefit, a cost, or no difference. Again, I found that participants were much more likely to express a belief in a relatedness benefit for location memory, rather than a cost, although a sizable minority (comparable to Experiment 4) did express a belief in a cost. Table 7 shows the proportion of participants in each group.

Table 7. Experiment 5a/5b: Proportions of participants who predicted a related benefit, related cost, or no difference

	Related Benefit	**Related Cost**	**No Difference**
Experiment 2a (*N*=400)	36.3%	20.3%	43.5%
Experiment 2b (*N*=336)	53.3%	12.2%	34.5%
Total (*N*=736)	37.9%	18.8%	43.3%

Discussion

In Experiments 5a and 5b, I again found that participants predicted that the locations of related items would be better remembered than those of unrelated items. A similar pattern emerged when I categorized individuals into those believing in a relatedness benefit, cost, or no difference, with more participants predicting a benefit of relatedness to location memory rather than a cost. The effect size obtained in Experiment 5a was smaller than previously observed, suggesting possible influences of description and/or scale order. However, the effect was found to be robust in Experiment 5b when the unrelated condition was described and rated first. In conclusion, participants' tendency to believe in a relatedness benefit for location memory appeared to be robust, and this effect was not attributable to bias arising from presentation order nor from any misunderstanding of the memory task.

Experiment 6

One interpretation of the behavior of the subset of participants who reported a benefit of relatedness in the location memory task is that they simply failed to consider task as a relevant attribute. That is, it might be the case that the naïve theories leading to *relatedness-as-benefit* are more accessible in the context of a given memory task whereas the theories leading to *relatedness-as-cost* are more likely to emerge when relatedness and task are considered in conjunction. In Experiments 4 and 5, participants made metacognitive judgments for a single task (in Experiment 4, either item or location memory; in Experiment 5, only location memory). That is, participants made their judgments in a kind of single evaluation mode (Hsee & Zhang, 2004; Hsee et al., 1999). While single evaluation involves participants making judgments about a singly presented option, joint evaluation induces participants to compare different alternatives against each other.

Hsee and colleagues (1999) argued that the juxtaposition of alternatives in a joint evaluation context leads to participants prioritizing evaluable attributes that can be directly compared, which in turn can result in surprising preference reversals compared to single evaluation. For example, Hsee (1996) found that in single evaluation, participants were willing to pay more for a dictionary with 10,000 entries in like-new condition compared to a dictionary with 20,000 entries with a torn cover, though the reverse was true in joint evaluation. Hsee and colleagues argued that in single evaluation, most participants would not know how to evaluate the desirability of a dictionary with 20,000 (or 10,000) entries, leading to them prioritizing the cover condition in their valuations (Hsee et al., 1999). On the other hand, the number of dictionary entries only became evaluable in joint evaluation, where alternatives were juxtaposed.

In Experiments 4 and 5, participants evaluated the effect of relatedness in joint evaluation (within-subjects) but evaluated the effect of task type in single evaluation (between-subjects). Relatedness has previously been reported to be a highly evaluable and salient attribute that influences judgments of learning even when it is manipulated between participants (e.g., Dunlosky & Matvey, 2001). However, other attributes have been reported to be less evaluable in a single evaluation context. For example, Koriat et al. (2004) reported that participants' recall predictions were not sensitive to retention interval when manipulated between-subjects, such that they predicted similar levels of recall in each condition (10 minutes vs. 1 day vs. 1 week). In contrast, a within-subjects manipulation resulted in a strong effect of retention interval in participants' recall predictions (Koriat et al., 2004). While participants failed to consider the influence of retention interval in a single evaluation context, it became evaluable and salient in a joint evaluation context, with Koriat and colleagues (2004) arguing that this attribute activated participants' beliefs about memory decline over time. Task type may be a similarly difficult-to-evaluate attribute, which could explain why the majority of my participants predicted a relatedness benefit even in the location memory task.

If the predicted relatedness benefit in the location memory task was driven by participants failing to consider how relatedness might operate across different task contexts, then this effect should be attenuated when task type is made more salient. In Experiment 6, I tested this hypothesis by eliciting participants' predictions in more of a joint evaluation context, such that each participant was asked about the influence of relatedness in both the item and location memory tasks. I anticipated that the within-subject manipulation of task would increase evaluability of this attribute, inducing participants to reduce the size of the predicted relatedness benefit for location memory compared to item memory. A key advantage of joint evaluation is that it allows us to assess to what degree participants are sensitive to the interaction.

Finally, another goal of Experiment 6 was to examine to what extent participants' beliefs about relatedness were similar across the two tasks. In Experiment 1, I proposed that the task vignettes led to the activation of distinct mechanistic theories about how relatedness influences memory (e.g., relatedness-as-benefit via creating associations between items; relatedness-as-cost via rendering individual items less distinct). In Experiments 1 and 2 participants made a single prediction about the effect of relatedness in one of the two tasks. In Experiment 3, participants will provide predictions for both tasks, thus allowing a consideration of their relation to each other within participants.

By examining the degree of agreement across participants' relatedness beliefs in the item and location tasks, we can assess to what extent they activated the same or different mechanistic pathways in the two tasks. Therefore, I pre-registered an analysis that examined the proportion of participants expressing task-concordant relatedness beliefs—in other words, predicting an effect of relatedness in the same direction for both tasks. For example, believing that relatedness would benefit both item and location memory would be concordant whereas believing that relatedness would benefit item memory but harm location memory would be discordant. Note that the "correct" belief would be discordant in this case because relatedness is beneficial in the item task and harmful in the location memory task. If the two task descriptions tend to activate the same mechanistic belief, then the proportion of participants expressing task-concordant beliefs would be higher than would be expected by chance—that is, participants would be more likely to express relatedness beliefs in the same direction. In a similar vein, though exploratory, I also correlated the magnitudes of the effects of relatedness across tasks.

Method

Participants

Data from 200 participants (118 women, 81 men, 1 other, *M* = 41.33 years, *SD* = 13.87) were analyzed. Participants were recruited from Prolific and compensated GB£0.38 for approximately three minutes of participation.

Materials and Procedure

The procedure was highly similar to Experiment 4, except that all participants were given a description of both the item memory task and the location memory task. The order of the two task descriptions was randomized for each participant. After reading both task descriptions, participants were asked to make predictions (for the related and unrelated item display) of their performance in each task. The four scales were presented on the same page; scale order (item vs. location first) was matched to task description order for each participant. After providing their predictions, participants were reminded of them on a new screen and asked to type in their reasoning, similar to Experiment 4.

Results

Data from three participants were not analyzed according to the exclusion criteria set in the pre-registration, leaving $N = 200$ after replacement.

Effects of relatedness and task on memory predictions

A 2 (Relatedness: Related vs. Unrelated Items) x 2 (Memory Type: Item vs. Location) within-subject ANOVA revealed a main effect of Relatedness, $F(1,199) = 35.10$, $MSE = 1.66$, $p < .001$, $\eta_G^2 = .018$, such that memory predictions were higher for related items than unrelated items. There was also a main effect of Memory Type, *F*(1,199) = 57.97, *MSE* = 2.46, *p* < .001, η_G^2 = .042, such that item memory predictions were higher than location memory predictions. Unlike Experiment 1, the interaction was not significant, *F*(1,199) = 1.02, *MSE* = 1.25, *p* = .314, η_G^2 < .001. Nevertheless, as pre-registered, I followed up with separate ANOVAs for item and location memory predictions. For item memory predictions, related items were predicted to be remembered better than unrelated items, *F*(1,199) = 26.15, *MSE* = 1.47, *p* < .001, η_G^2 = .024. For location memory predictions, related item locations were also predicted to be remembered better than unrelated item locations, *F*(1,199) = 14.63, *MSE* = 1.45, *p* < .001, η_G^2 = .012. Table 8 shows the mean memory predictions in each condition.

Table 8. Experiment 6: Mean memory predictions (SD) by item relatedness and memory type

	Related	**Unrelated**
Item Memory	5.47 (2.10)	4.85 (1.86)
Location Memory	4.54 (2.21)	4.08 (1.90)

Effects of task on relatedness belief direction

Based on their related vs. unrelated prediction differences, I sorted participants into nine groups: those that predicted a relatedness benefit, cost, or no difference for item memory, crossed with predicting a relatedness benefit, cost, or no difference for location memory. Table 9 shows the proportion of participants in each group. I pre-registered an analysis comparing the size of the predicted benefit for the subset of participants who predicted a relatedness benefit for both memory types (item and location; 29.5% of all participants). For these participants, the size of the predicted benefit did not differ for item memory (*M* = 1.73, *SD* = 1.62) vs. location memory (*M* = 1.58, *SD* = 1.58), *t*(58) = 1.42, *p* = .162, d_z = 0.18 [95% CI: -0.07, 0.44], mirroring the result from Experiment 4.

Table 9. Experiment 6: Proportions of participants who predicted a related benefit, related cost, or no difference for the item vs. location memory tasks

	Item		
Location	Relatedness Benefit	Relatedness Cost	No Difference
Relatedness Benefit	29.5%	2.0%	7.0%
Relatedness Cost	6.5%	2.0%	3.5%
No Difference	13.0%	4.0%	32.5%

Belief concordance across item and location tasks

I pre-registered an analysis examining whether participants were more likely to espouse task-concordant relatedness beliefs (i.e., predicting a relatedness effect in the same direction for both tasks; the three diagonal cells in Table 9) compared to discordant beliefs (i.e., the other six cells in Table 9). A chi-square goodness of fit test found that the proportion of task-concordant relatedness beliefs (64.5%) was significantly different than would be expected by chance (3 out of 9 cells, or 33.3%), $\chi^2(1) = 87.42$, *p* < .001, suggesting that participants' relatedness beliefs tended to be consistent across the item and location memory tasks. This pattern echoes the results of the ANOVA: By and large, most participants' predictions were not sensitive to an interaction between task and relatedness.

As another measure of belief consistency across tasks, I conducted an exploratory analysis on the correlation between participants' item and location relatedness predictions (their related-unrelated prediction differences). Consistent with the above analysis, I found a positive correlation across the two tasks, *r*(198) = .14, *p* = .049.

Effects of relatedness and task on self-reported prediction reasoning

Since participants were given one text box to explain their reasoning for both tasks, I found that some participants reasoned only about relatedness or only about task. This made it difficult to follow Experiment 4's scheme where participants only had to explain their beliefs about relatedness. Instead, I devised a simplified scheme wherein participants' responses were coded for how sensitive they were to each potential effect: (1) a main effect of task (e.g., *I think would have an easier time remembering the words but would struggle to remember exactly where they were*); (2) a main effect of relatedness (e.g., *I think would be slightly easier if they are from the same category*); (3) an interaction between task and relatedness (e.g., *I think it is easier to remember words within the same category because I can group them together, but I think remembering the location of words in different categories will be easier as there is more of a uniqueness to them which may help me identify the locations in memory*).

As in Experiment 4, I had two independent coders blind to participants' predictions and condition assignment, and each reason could be assigned to multiple categories under the coding scheme. Cohen's κ ranged from 0.84 to 0.97, suggesting almost perfect agreement between the two raters (Cohen, 1960). Table 10 shows the proportions of participants who described each effect in their self-reported reasoning. Of note, only a small minority of participants (11 out of 200, 5.5%) were able to accurately anticipate an interaction between task and relatedness.

Table 10. Experiment 6: Percentages of participants anticipating each effect in their self-reported reasoning

	Task Concordance **(*N*=128)**	**Task Disconcordance** **(*N*=72)**
Main Effect of Task	39.1%	36.1%
Main Effect of Relatedness	27.3%	27.8%
Task x Relatedness Interaction	3.1%	9.7%

Discussion

In Experiment 6, overall, participants predicted a similar-sized benefit of relatedness for the location memory task and the item memory task. Their predictions demonstrated both a main effect of relatedness and a main effect of task: They predicted that memory would be better for related items than for unrelated items and that item memory would be worse than location memory (Lu et al., 2023). They were not, however, sensitive to the interaction between the two factors—that relatedness would operate differently in these two memory tasks (Lu et al., 2023). These results depart from my predictions and from Experiment 4, where I had found an interaction of task and relatedness such that participants' aggregate predictions accurately captured a larger benefit of relatedness for item memory than for location memory. While Experiment 4 manipulated task between-subjects (i.e., task was considered in a single evaluation context), Experiment 6 manipulated task within-subject (i.e., task was presented in a joint evaluation context). I had initially predicted that making task more salient in a joint evaluation context (where participants are exposed to both tasks) might reduce the relatedness benefit in the location memory task. Contrary to this expectation, I found a null interaction between task and relatedness such that participants predicted an equally large benefit of relatedness in the item and location memory tasks.

In Experiment 4, I noted that the aggregate can be misleading, and that the task by relatedness interaction was driven by more participants predicting a benefit for item memory and a cost for location memory, rather than, for example, participants predicting a relatedness benefit to a lesser extent in the location task. In Experiment 6, I found that almost no participants actually predicted this interaction. Instead, participants who predicted a benefit predicted that the effect would be the same size across tasks (similar to Experiment 4). These results would appear to refute an account of the relatedness benefit in the location task of Experiments 4 and 5 based on the idea that participants were simply failing to consider task as a relevant attribute. Not only did participants continue to predict a benefit of relatedness in the location memory context here, but the predicted relatedness effects for the two tasks converged such that there was no difference between the item and location relatedness predictions. That is, rather than inducing *task-divergence*, exposing participants to both tasks in a joint evaluation context appeared to result in high *task-concordance*: 64.5% of participants' relatedness beliefs exhibited task-concordance (i.e., predicting an effect of relatedness in the same direction for the item and location tasks), which was far higher than would be expected by chance. Furthermore, the joint evaluation context appeared to result in even fewer people predicting a relatedness cost in the location memory task relative to the single evaluation context (12.5% in Experiment 6, vs. 19.5% in Experiment 4 and 18.8% in Experiment 5a/b). Thus, the joint evaluation of task in Experiment 6 may have increased the likelihood that a theory of relatedness-as-benefit was retrieved and applied in both tasks.

Contrary to my hypotheses, the single evaluation context might be necessary for the retrieval of the less accessible relatedness-as-cost idea. This would be consistent with the claim that the relatedness-as-benefit mechanism represents the stronger mental attractor. When a vignette (i.e., item memory) that clearly primes the relatedness-as-benefit mechanism is presented in proximity to a vignette (i.e., location memory) that sometimes primes the relatedness-as-cost mechanism, as in Experiment 6, the latter simply loses the competition.

Experiment 7

Given the lack of a task by relatedness interaction observed in Experiment 6, I conducted Experiment 7 as a close replication of Experiment 4 to determine whether the interaction is indeed robust when task is manipulated between-subjects. Thus, Experiment 7's procedure was highly similar to Experiment 4 with some changes. Importantly, I changed the wording of the prompt that asked participants to explain their predictions: Instead of simply reminding them of the numeric value of their predictions, I directed them to explain why their predictions about relatedness were in a specific direction (i.e., "you predicted that you would remember more/less/the same number of items from the related/unrelated item display"). This was to make it clear to participants that they should reason about the effects of relatedness in their descriptions.

Another goal of Experiment 7 was to obtain a measure of participants' subjective confidence in their relatedness predictions. While two participants might express similar beliefs about the effect of relatedness, they could nevertheless differ in the degree of confidence or *credence* in these beliefs. That is, credence is a measure of how strongly or weakly one holds a particular belief (e.g., one is almost certain that the sun will rise tomorrow but likely less certain that it will rain). There is some debate in the philosophical literature as to what extent beliefs and credences might be independent or reduce to each other (Jackson, 2022). According to the self-consistency model of subjective confidence (Koriat, 2012; see also Koriat, 2013; 2018; 2024), when faced with a question, we retrieve a variety of information from memory that is relevant to the question. The level of confidence that one has in their answer is determined by the extent to which the retrieved representations consistently support the chosen answer: If most of the retrieved information points toward the same answer, this is expressed as high confidence; conversely, if the information is contradictory or supports different answers, then confidence will be lower.

Confidence has been found to track both the consistency of an answer within an individual as well as the consensus regarding the answer across individuals (Koriat, 2013; 2018; 2024). Thus, applying this idea to metacognitive predictions about relatedness, credence can be taken as a rough measure of how likely participants will traverse the same mental pathway when thinking about relatedness. I hypothesized that there would be task differences in participants' belief credence: Because the relatedness benefit is more readily expressed in the item task, participants may also be more confident that this belief is correct for the item task.

Method

Participants

Data from 400 participants (175 women, 218 men, 3 other, 4 unknown, *M* = 37.91 years, *SD* = 12.65) were analyzed. Participants were recruited from Prolific and compensated GB£0.38 for approximately three minutes of participation.

Materials and Procedure

As in Experiment 4, participants were randomly assigned to the item memory group (*N* = 195) or the location memory group (*N* = 205). After reading the experiment vignette, participants predicted their memory performance from 0 to 10 items for the related and unrelated conditions.

Belief Confidence. After making their predictions, participants were shown the following prompt and asked to rate their confidence from 1 (Not at all confident) to 7 (Extremely confident):

> “In the last task we asked you to make predictions about the number of words you would be able to correctly [recall/recall the location]. You estimated that you would remember [more items/the same number of items] in the display where [all items are from the same category/all items are from different categories], [compared to/and] the display where [all items are from the same category/all items are from different categories]. How confident are you that this prediction is correct?”

Self-Reported Reasoning. On the next screen, participants were shown the same prompt as the previous one that reiterated both the nature of the memory predictions (item or location) as well the direction of their relatedness predictions. The final sentence was changed to “Please tell us about your reason(s) for this prediction”, and participants were to type their responses into an on-screen text field.

Results

In accordance with the exclusion criteria set in the pre-registration, data from 11 participants were not analyzed, leaving *N* = 400 after replacement.

Effects of relatedness and task on memory predictions

A 2 (Relatedness: Related vs. Unrelated Items) x 2 (Memory Type: Item vs. Location) mixed-subjects ANOVA revealed a main effect of Relatedness, $F(1,398) = 57.68$, $MSE = 1.62$, $p < .001$, $\eta_G^2 = .034$, such that memory predictions were higher for related items than for unrelated items. There was also a main effect of Memory Type, $F(1,398) = 6.29$, $MSE = 5.06$, $p = .013$, $\eta_G^2 = .012$, such that item memory predictions were higher than location memory predictions. As in Experiment 4, the interaction was significant, $F(1,398) = 4.31$, $MSE = 1.62$, $p = .039$, $\eta_G^2 = .003$. For item memory predictions, related items were predicted to be remembered better than unrelated items, $F(1,194) = 39.83$, $MSE = 1.86$, $p < .001$, $\eta_G^2 = .055$. For location memory predictions, related item locations were also predicted to be remembered better than unrelated item locations, though to a lesser degree, $F(1,204) = 18.12$, $MSE = 1.40$, $p < .001$, $\eta_G^2 = .018$. Table 11 shows the mean memory predictions in each condition.

Table 11. Experiment 7: Mean memory predictions (SD) by item relatedness and memory type

	Related	**Unrelated**
Item Memory (*N*=195)	5.71 (1.83)	4.84 (1.81)
Location Memory (*N*=205)	5.13 (1.90)	4.63 (1.77)

Effects of task on relatedness belief direction

As in Experiment 4, I sorted participants into three groups according to their related vs. unrelated prediction differences: (1) those who predicted that they would remember more items/item locations for the related items (related benefit), (2) those who predicted that they would remember fewer items/item locations for the related items (related cost), and (3) those who predicted that they would remember equal numbers for related vs. unrelated items/item locations (no difference). Unlike Experiment 4, a chi-square test suggested that the proportions in each group did not differ significantly for location vs. item memory, $\chi^2(2) = 4.53$, $p = .104$. Table 12 shows the proportion of participants in each group.

Table 12. Experiment 7: Proportions of participants who predicted a related benefit, related cost, or no difference for item vs. location memory

	Related Benefit	**Related Cost**	**No Difference**
Item Memory **(*N*=195)**	55.9%	9.2%	34.9%
Location Memory **(*N*=205)**	45.9%	13.7%	40.5%

I conducted an exploratory analysis on the combined count data from Experiment 4 and Experiment 7 to obtain more power to detect any effect of task on participants' belief type. A chi-square test suggested that the proportions in each group differed for location vs. item memory, $\chi^2(2) = 21.66$, $p < .001$. Pairwise comparisons with Bonferroni-adjusted *p*-values revealed that significantly more participants predicted a benefit for item memory compared to location memory, $\chi^2(1) = 12.99$, $p < .001$, and significantly more participants predicted a cost for location memory compared to item memory, $\chi^2(1) = 15.17$, $p < .001$, whereas the proportion of participants predicting no difference did not differ by task, $\chi^2(1) = 1.06$, $p = .303$.

Effects of task on relatedness predictions for each belief direction

I conducted pre-registered analyses that compared the magnitude of participants' relatedness beliefs (for item and location) within each type of expressed belief (i.e., benefit vs. cost). For the subset of participants who predicted a relatedness benefit, the size of the predicted benefit did not differ for item vs. location memory, *t*(193.81) = 1.69, *p* = .092. A similar null effect was found for the subset of participants who predicted a relatedness cost, *t*(31.08) = 0.75, *p* = .461, although this result should be interpreted with caution given the low *Ns*. Table 12 shows the magnitude of the predicted relatedness effects within each category.

I conducted an exploratory analysis on the combined data from Experiment 4 and Experiment 7 to obtain more power to detect any task differences for the predicted effects within each subset. Again, the size of the predicted benefit was found to be the same across task, *t*(393.24) = 0.58, *p* = .565, as was the size of the predicted cost, *t*(42.86) = 1.48, *p* = .147.

Effects of task on relatedness belief confidence

I found a similar level of confidence in participants' relatedness predictions for the item (*M* = 4.47) and location (*M* = 4.34) groups. This result was obtained in both a non-parametric rank test (pre-registered), *W* = 20858, *p* = .437, and a parametric test (exploratory), *t*(393.17) = 0.97, *p* = .333. When I conducted these pre-registered comparisons separately for each category of expressed belief (i.e., benefit vs. cost vs. no difference), I found no difference between the item and location groups (all *p*s > .194). Table 13 shows prediction confidence within each category.

Table 13. Experiment 4: Participants' mean (SD) related vs. unrelated prediction magnitudes and prediction confidence

	Item Memory (*N*=195)			**Location Memory (*N*=205)**		
	Related Benefit (*N* = 109)	Related Cost (*N*=18)	No Difference (*N*=68)	Related Benefit (*N*=94)	Related Cost (*N*=28)	No Difference (*N*=83)
Related-Unrelated Difference	2.02 (1.29)	-2.78 (2.51)	0	1.76 (0.91)	-2.25 (2.05)	0
Prediction Confidence	4.75 (1.02)	3.94 (1.39)	4.15 (1.26)	4.56 (1.24)	4.21 (1.17)	4.13 (1.59)

Based on Table 13, I wondered whether participants' confidence might differ depending on the type of belief that they expressed (related benefit, cost, or no difference). An exploratory 3 (Belief Direction: Benefit vs. Cost vs. No Difference) x 2 (Memory Type: Item vs. Location) between-subjects ANOVA revealed a main effect of Belief Direction, $F(2,394) = 8.65$, $MSE = 1.62$, $p < .001$, $\eta_G^2 = .042$. Echoing the earlier analyses, there was no main effect of Memory Type, $F(1,394) = 0.02$, $MSE = 1.62$, $p = .888$, $\eta_G^2 < .001$, nor did these two factors interact, *F*(2,394) = 0.64, *MSE* = 1.62, *p* = .526, η_G^2 = .003. Post-hoc analyses (Tukey-adjusted) indicated that participants who predicted a relatedness benefit expressed higher levels of confidence compared both to participants predicting a cost, *t*(394) = 2.73, *p* = .018, and to the participants predicting no difference, *t*(394) = 3.78, *p* < .001. Participants who predicted a cost of relatedness or no difference expressed similar levels of confidence, *t*(394) = 0.28, *p* = .959. An exploratory correlational analysis further revealed that participants' belief credence was positively correlated with their related-unrelated predictions, *r*(398) = .18, *p* < .001.

Figure 8 shows the correlation between participants' relatedness predictions and the confidence that their prediction was in the correct direction. Since participants were more confident when predicting a benefit than a cost or no difference, I wondered whether this effect was driving the overall correlation. Thus, I conducted separate exploratory correlations within each belief direction subset. For participants predicting a benefit, their belief credence was positively correlated with the magnitude of that benefit, *r*(201) = .19, *p* = .007. For participants predicting a cost, their belief credence trended non-significantly with the magnitude of that cost, *r*(44) = -.30, *p* = .764.

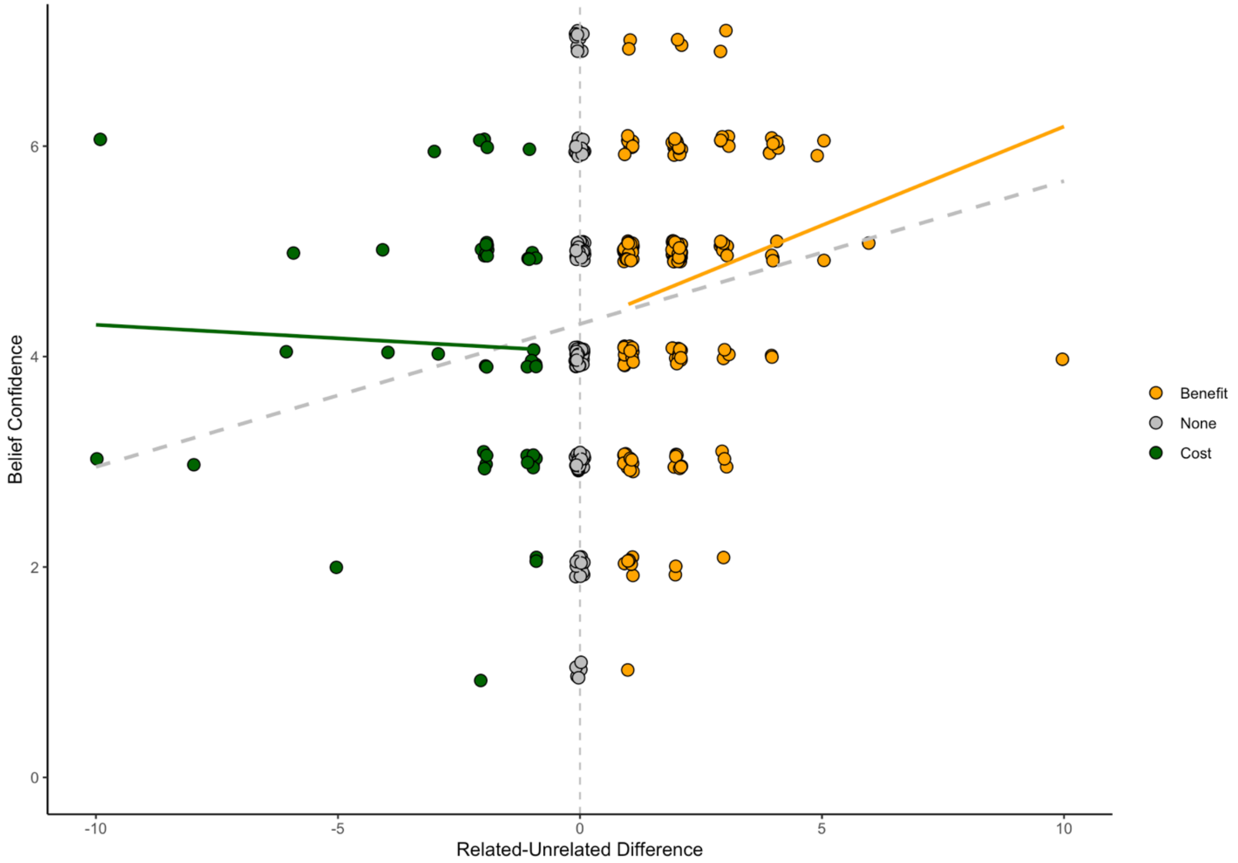


Figure 8. Correlation plot showing the relation between the magnitude of participants' relatedness predictions and their belief confidence in Experiment 7. Solid lines represent the correlation within each belief direction subset (green line for benefit, orange line for cost); dashed line represents overall correlation. Circles in the background represent individual participants' predictions (green circles represent participants who predicted a relatedness benefit; orange circles represent participants who predicted a cost; grey circles represent participants who gave equal predictions both conditions).

Effects of task and relatedness on self-reported prediction reasoning

Participants' self-reported reasons for their predictions were classified using the same scheme from Experiment 4. Coders again were blind to participants' predictions and condition assignment. The ‘memory ability’ and ‘other reasons’ category showed moderate inter-rater reliability (κ = 0.60 and .52 respectively). For the other categories, κ ranged from 0.83 to 0.92, suggesting almost perfect agreement between the two raters (Cohen, 1960). Table 14 shows the percentage of participants expressing a particular type of reason (as coded by the primary coder).

Table 14. Experiment 7: Percentages of participants’ self-reported reasoning behind their predictions. Note that responses could be classified into multiple reasons, and some were unclassifiable/unusable, hence columns do not sum to 100%

	Item Memory (*N*=195)			**Location Memory (*N*=205)**		
	Related Benefit	Related Cost	No Difference	Related Benefit	Related Cost	No Difference
Imagery	7.3%	-	-	9.6%	3.6%	-
Association	47.7%	5.6%	5.9%	29.8%	10.7%	3.6%
Distinctiveness	-	22.2%	1.5%	2.1%	46.4%	1.2%
Unexplained Related Preference	17.4%	-	-	34.0%	-	-
Unexplained Unrelated Preference	-	27.8%	-	-	21.4%	-
Indifference	-	11.1%	39.7%	-	-	41.0%
Memory Ability	3.7%	11.1%	20.6%	6.4%	10.7%	25.3%
Other	20.2%	11.1%	2.9%	14.9%	3.6%	1.2%

Discussion

Experiment 7 largely replicated the results of Experiment 4: Participants predicted that related items would be better remembered than unrelated items for both the item memory and location memory tasks, although the predicted relatedness benefit was smaller in the location memory task. As in Experiment 4, participants' aggregate predictions exhibited a task by relatedness interaction. These findings contrast with those of Experiment 6: When task was manipulated between-subjects (Experiment 4, Experiment 7), the overall predicted benefit was larger for item compared to location, but when task was manipulated within-subject (Experiment 6), the predicted benefit was similar for both tasks. Although Experiment 7 did not statistically replicate the effect of task on belief direction, a combined analysis across Experiments 4 and 7 was consistent with the notion that the task by relatedness interaction was driven by more people predicting a benefit for item and a cost for location (task had no effect on the proportion of people predicting no difference). Within each relatedness belief subset (benefit and cost), the magnitude of the predicted effect was the same across task.

Participants' self-reported reasoning data were similar to those of Experiment 4: People predicting a relatedness benefit, they most often invoked the idea of making more connections/associations between the related items, whereas people predicting a relatedness cost were more likely to invoke the notion of reduced distinctiveness or increased confusability. Table 14 further suggests that more participants were able to articulate the associations idea for the item memory task compared to the location memory task, and more participants were able to articulate the distinctiveness idea for the location memory task compared to the item memory task, with a proportionate reduction in those expressing an "unexplained related/unrelated preference".

An additional goal of Experiment 7 was to examine participants' belief credence, as measured by subjective confidence in their predictions about relatedness. Contrary to my hypothesis, participants did not express higher confidence in their relatedness beliefs for the item task. However, participants who predicted a benefit were more confident in this belief than were participants who predicted a cost or no difference, irrespective of task. Since the effect of task on belief direction did not reach significance in Experiment 7, this could explain why I did not find an effect of task on confidence. Nevertheless, the difference in credence across the groups that predicted an effect of relatedness in different directions is notable. It is possible that the lower confidence in the no difference group reflects a mixture of participants with a strong commitment to their being no difference and a more apathetic group of participants with little commitment either way.

If one takes confidence to be a measure of the degree of consistency across a given participant's retrieved representations (Koriat, 2012), then the higher credence expressed in the relatedness-as-benefit group suggests that this retrieval path is more consistently activated in the current context. The higher credence associated with the relatedness-as-benefit idea is therefore consistent with the argument that I have made throughout that this naïve theory appears to be the stronger mental attractor. Interestingly, while the task description (item vs. location) again influenced the proportion of participants who ended up retrieving relatedness-as-benefit or relatedness-as-cost ideas, once retrieved, the credence associated with each belief was the same.

Experiment 8

In the previous experiments in this chapter, participants predicted a benefit of relatedness overall in the aggregate, but they varied in terms of both their predictions about relatedness as well as the reasoning that they used to support these predictions. According to participants' self-reported reasoning, people who thought that relatedness benefits memory tended to invoke explanations along the lines of relatedness promoting *associations and connections* whereas people who thought that relatedness hinders memory tended to invoke explanations pertaining to relatedness increasing *confusability* and reducing *distinctiveness*. Based on these findings, I have advanced the hypothesis that the differential activation of various naïve theories of memory was driving the differences in participants' relatedness beliefs. In this view, participants' written explanations offer insight into their naïve theories that led them to conclude a particular effect of relatedness. An alternative view, however, is that these explanations were merely post-hoc justifications. That is, participants might have expressed their predictions about relatedness without much insight at the time, but then subsequently searched for further supporting information only when asked to justify them. Thus, I investigated whether activating different memory mechanisms *prior* to eliciting participants' relatedness predictions would lead to downstream differences in their predictions.

In Experiment 8, I primed participants with two distinct narratives about memory (associations vs. distinctiveness). Previously, narrative prime manipulations have been successfully used to temporarily induce beliefs; for example, Miele and Molden (2010) had participants first read an article that was designed to induce them to believe that intelligence was either fixed or malleable, which then led to differences in how they interpreted encoding fluency. Participants who were led to believe that intelligence was fixed reported lower levels of comprehension for a low-fluency (vs. high-fluency) text, suggesting that they took low fluency to indicate that they were reaching the limit of their abilities. On the other hand, participants who were led to believe that intelligence was malleable did not show this association between fluency and comprehension.

The priming manipulation was as follows: Before reading the experimental vignette and making predictions, half of the participants read a passage about how making associations between items is key to memory whereas the other half read a passage about how distinctiveness between items is key to memory. Based on the naïve theory framework, I predicted that participants who read the associations prime would predict a stronger benefit of relatedness than would participants who read the distinctiveness prime, and further anticipated that participants would invoke the primed concept when asked to explain their relatedness predictions.

The experimental procedure of Experiment 8 was similar to the previous experiments (location memory vignette only), with some refinements to the experimental procedure. As in Experiment 7, I prompted participants to explain why their predictions about relatedness were in a particular direction, to make it clear that they should reason about the effect of relatedness specifically. Based on the results of Experiments 4 and 7, I also asked participants to *self-categorize* their written explanations as one of several multiple-choice responses (i.e., associations; distinctiveness; mental imagery; other) that best described their explanation. I anticipated that this would make it easier for participants to express themselves, as they would only have to recognize the concept that matched their thoughts rather than articulate it. I also hoped that this would facilitate subsequent coding and data analysis.

Method

Participants

Data from 400 participants (175 women, 218 men, 3 other, 4 unknown, *M* = 37.91 years, *SD* = 12.65) were analyzed. Participants were recruited from Prolific and compensated GB£0.50 for approximately four minutes of participation.

Materials and Procedure

Prime Passage. Participants were randomly assigned to read either the associations prime or the distinctiveness prime (see Appendix B for the full prime passages); they were told that the passage explained a basic tenet about human memory. The prime passages did not mention relatedness or location memory.

Comprehension Check and Summary. After participants indicated that they had finished reading the passage, they answered a multiple-choice comprehension check question: "Researchers believe that ______ is the key to memory: (i) [making associations/distinctiveness], (ii) rehearsal, (iii) mental imagery". The order of the three options were randomized for each participant. On the next screen, they were asked to summarize the point of the passage that they had just read in one sentence. The comprehension check and summary tasks were designed to enhance prime effectiveness.

Related and Unrelated Memory Prediction. Next, as in previous experiments, participants were given a description of the location memory task and asked to make predictions for the related and unrelated item displays.

Self-Reported Reasoning. After making their predictions, participants were shown the prompt from Experiment 7 and asked to type their reasoning into an on-screen text field.

Following the open-ended response, participants were asked "Which of the following best describes the reasoning behind your predictions?" and asked to select one of multiple options. The options that each participant saw were customized according to their relatedness predictions. If they predicted a relatedness benefit, the first three options were "Easier to form associations/connections between items from the same category", "Items from the same category are more distinctive and hence memorable", "Can form better mental imagery for items from the same category". If they predicted a relatedness cost, the first three options were "Easier to form associations/connections between items from the different categories", "Items from the different categories are more distinctive and hence memorable", "Can form better mental imagery for items from the different categories". If they predicted the same number for the related and unrelated conditions, the first three options were "Equally easy to form associations/connections between items from the same category and different categories", "Items from the same category and different categories equally distinctive and hence memorable", "Can form similar mental imagery for items from the same category and different categories". The order of the first three options was randomized for each participant. The last two options were always "I did not think there was any reason the two displays would differ in terms of my ability to remember the locations" and "My reason is not listed (explain on next page)", the latter allowing participants to write their own reasoning.

Results

I used the following pre-registered exclusion criteria: (1) did not give both related and unrelated predictions; (2) spent less than 25 seconds[4] reading the prime passage; (3) did not answer the prime comprehension check question correctly; (4) indicated that they were not paying attention or did not give effort during the task (self-report at the end of the study). Based on these criteria, data from 95 participants were not analyzed, leaving *N* = 400 after replacement. One participant who participated after the stopping rule was also excluded. The final sample sizes in each group were *N* = 197 in the associations prime condition and *N* = 203 in the distinctiveness prime condition.

[4] I originally pre-registered a 30-s cutoff, however, this rule proved to be too strict and led to too much data loss. I analyzed the data at both the 25-s cutoff and 30-s cutoff and obtained the same pattern of results.

Effects of relatedness and prime on memory predictions

A 2 (Relatedness: Related vs. Unrelated Items) x 2 (Prime Type: Associations vs. Distinctiveness) mixed ANOVA revealed a main effect of Relatedness such that memory predictions were higher for related items than unrelated items, $F(1,398) = 22.23$, $MSE = 1.32$, $p < .001$, $\eta_G^2 = .011$, and no main effect of Prime Type, $F(1,398) = 0.93$, $MSE = 5.47$, $p = .334$, $\eta_G^2 = .002$. More importantly, these were qualified by a significant interaction, $F(1,398) = 41.13$, $MSE = 1.32$, $p < .001$, $\eta_G^2 = .020$. Follow-up analyses revealed that participants who read the associations prime predicted a significant relatedness benefit, $F(1,196) = 70.81$, $MSE = 1.14$, $p < .001$, $\eta_G^2 = .053$, whereas participants who read the distinctiveness prime predicted a nonsignificant effect in the opposite direction, $F(1,202) = 1.29$, $MSE = 1.50$, $p = .257$, $\eta_G^2 = .002$. Table 15 shows the mean memory predictions in each condition.

Table 15. Experiment 8: Mean memory predictions (SD) by item relatedness and prime type

	Related	**Unrelated**	**Difference**
Associations Prime (*N*=197)	5.21 (1.92)	4.30 (1.89)	0.91
Distinctiveness Prime (*N*=203)	4.85 (1.78)	4.99 (1.77)	-0.14

Effects of prime on relatedness belief direction

In the following analyses, I sorted participants according to their relatedness predictions (expressed belief in a relatedness benefit, cost, or no difference), and the type of reasoning that they used to support these predictions (associations, distinctiveness, imagery/other[5]), as well as how these were influenced by the prime that they read. I pre-registered two chi-square analyses: (1) whether the two primes were associated with different kinds of relatedness beliefs; (2) whether the two primes were associated with different types of reasoning to support those beliefs. I also report an exploratory chi-square examining whether different relatedness beliefs were associated with different types of supportive reasoning.[6]

Primes and Relatedness Beliefs. A pre-registered chi-square test showed that prime type had a significant influence on the proportions of participants expressing each type of belief, $\chi^2(2) = 48.55$, $p < .001$. Participants who read the associations prime overwhelmingly tended to express a belief in a relatedness benefit compared to a cost whereas participants who read the distinctiveness prime were roughly equal in expressing a belief in a benefit or a cost. Pairwise comparisons (Bonferroni-adjusted) revealed that more participants predicted a benefit if they read the associations prime compared to the distinctiveness prime, $\chi^2(1) = 32.58$, $p < .001$, but fewer participants predicted a cost, $\chi^2(1) = 36.48$, $p < .001$, while the proportion predicting no difference were equally split, $\chi^2(1) = 0.28$, $p = .597$. Exploratory analyses showed that for the subset of participants who predicted a relatedness benefit, the size of the predicted benefit did not differ for the associations prime ($M = 1.87$, $SD = 0.91$, $N = 115$) vs. the distinctiveness prime ($M = 1.85$, $SD = 0.94$, $N = 60$), $t(117.22) = 0.13$, $p = .895$; similarly, for the subset of participants who predicted a relatedness cost, the size of the predicted cost did not differ for the associations prime ($M = -1.95$, $SD = 1.43$, $N = 19$) vs. the distinctiveness prime ($M = -1.93$, $SD = 1.08$, $N =$

[5] These two categories were combined for analysis as pre-registered. No participants selected “I did not think there was any reason the two displays would differ in terms of my ability to remember the locations”.

[6] The count data form a three-way contingency table (2 x 3 x 3). An exploratory log-linear analysis (Field et al., 2012) showed that the three-way interaction was not significant ($p = .091$): that is, the conditional relation between any pair of variables given the third one was the same at each level of the third variable. Thus, the data can be appropriately understood by interpreting the three two-way tables.

72), *t*(23.66) = 0.05, *p* = .962. Table 16 shows the proportion of participants in each group in Experiment 8.

Table 16. Experiment 8: Proportions of participants who predicted a related benefit, related cost, or no difference as a function of prime type

	Relatedness Benefit	**Relatedness Cost**	**No Difference**
Associations Prime (*N* = 197)	58.4%	9.6%	32.0%
Distinctiveness Prime (*N* = 203)	29.6%	35.5%	35.0%

Primes and Self-Reported Reasoning. I examined whether prime type changed how participants explained their relatedness predictions in the self-reported reason multiple choice question (i.e., associations, distinctiveness, imagery/other). A pre-registered chi-square test showed that prime type had a significant influence on how participants reasoned about their predictions, $\chi^2(2) = 44.33$, *p* < .001. Pairwise comparisons (Bonferroni-adjusted) revealed that more participants used association-based reasoning if they read the associations prime compared to the distinctiveness prime, $\chi^2(1) = 25.35$, *p* < .001, whereas more participants used distinctiveness-based reasoning if they read the distinctiveness prime compared to the associations prime, $\chi^2(1) = 40.23$, *p* < .001. The proportion of participants using imagery or other reasoning did not differ across the primes, $\chi^2(1) = 1.19$, *p* = .275. Table 17 shows the proportion of participants in each group.

Table 17. Experiment 8: Proportions of participants who explained their predictions via associations, distinctiveness, or other reasoning as a function of prime type

	Associations Reasoning	**Distinctiveness Reasoning**	**Imagery or Other Reasoning**
Associations Prime (*N* = 197)	53.3%	18.3%	28.4%
Distinctiveness Prime (*N*=203)	28.1%	48.8%	23.2%

Relatedness Beliefs and Self-Reported Reasoning. I explored whether participants who predicted a relatedness benefit, cost, or no difference tended to rely on different kinds of reasoning to support their predictions (i.e., associations, distinctiveness, imagery/other). An exploratory chi-square test showed that belief direction had a significant influence on how participants reasoned about their predictions, $\chi^2(2) = 159.44$, $p < .001$. The majority of participants who expressed a benefit of relatedness tended to support their predictions via an associations-based account (i.e., easier to form associations between related items); in contrast, the majority of participants who expressed a cost of relatedness tended to do so via a distinctiveness-based account (i.e., unrelated items are more distinctive). To simplify the follow-up analyses, I conducted pairwise comparisons (Bonferroni-adjusted) on whether the direction of relatedness beliefs was associated with the proportion of associations vs. distinctiveness reasoning (disregarding the other reasoning column). Participants who predicted a benefit were more likely to use association-based reasoning over distinctiveness-based reasoning, $\chi^2(1) = 102.66$, $p < .001$; participants who predicted a cost were more likely to use distinctiveness-based reasoning over association-based reasoning, $\chi^2(1) = 91.11$, $p < .001$. Participants who predicted no difference did not differ with respect to these two kinds of reasoning, $\chi^2(1) = 3.09$, $p = .079$. Table 18 shows the proportion of participants in each group.

Table 18. Experiment 8: Proportions of participants who explained their predictions via associations, distinctiveness, or other reasoning as a function of relatedness belief type

	Associations Reasoning	**Distinctiveness Reasoning**	**Imagery or Other Reasoning**
Relatedness Benefit (*N* = 175)	67.4%	18.8%	29.1%
Relatedness Cost (*N* = 91)	3.4%	79.1%	9.7%
No Difference (*N* = 134)	21.7%	49.5%	38.1%

Self-Reported Reasoning Categorization. As an additional check on participants' self-categorizations, I had two independent coders code participants' self-reported reasoning following the scheme from Experiment 4. Two additional categories were added: if they relied on their prior experiences (e.g., *I know from previous experiences that recalling things that stand out tends to be easier than a list of similar items*), and if they mentioned the information presented in the prime explicitly (e.g., *I based my judgments on the first excerpt I were told to read regarding memory and distinctive information*). Coders were blind to participants' predictions, condition assignment, and self-categorizations. The 'other reasons' category exhibited much lower inter-rater reliability (κ = 0.26), and hence was omitted. For the other categories, κ ranged from 0.63 to 0.87, suggesting substantial to almost perfect agreement between the two raters (Cohen, 1960). Among the participants who self-classified their reasoning as falling into either *associations* or *distinctiveness,* this classification was also given by the primary coder over 40% of the time (a plurality), suggesting that some participants had difficulty fully expressing these ideas in their free reports. Table 18 shows the percentage of participants expressing a particular type of reason (as coded by the primary coder), compared against participants' self-categorizations.

Table 19. Experiment 8: Percentages of participants' self-categorized reasoning compared to experimenter-coded

	Self-Categorized Reasoning		
	Associations **(*N* = 162)**	**Distinctiveness** **(*N* = 135)**	**Imagery or Other** **(*N* = 103)**
Imagery	1.2%	1.5%	10.7%
Association	46.9%	8.9%	15.5%
Distinctiveness	3.7%	43.7%	6.8%
Unexplained Related Preference	14.8%	3.7%	4.9%
Unexplained Unrelated Preference	-	5.9%	1.0%
Indifference	13.0%	19.3%	19.4%
Memory Ability	7.4%	11.1%	26.2%
Past Experience	2.5%	1.5%	3.9%
Prime Information	7.4%	12.6%	-

Discussion

Consistent with earlier experiments, in Experiment 8 I found that participants predicted a relatedness benefit for location memory overall. Critically, the priming manipulation had a significant effect: Participants who read the associations prime predicted a significant relatedness benefit in the aggregate whereas participants who read the distinctiveness prime predicted a nonsignificant effect in the opposite direction. Further analyses revealed that participants were more likely to predict a benefit and less likely to predict a cost when they had read the associations prime versus the distinctiveness prime; the proportion of participants predicting no difference was not moderated by prime. Within each subset of participants who predicted a relatedness benefit or cost, the magnitude of the predicted benefit or cost did not differ across primes.

Overall, priming participants with different narratives about memory led to them expressing different beliefs about how relatedness would influence memory in a location memory task. It is critical to note that the primes contained no mention of relatedness or location memory: that is, participants could not have taken any ideas about the effects of relatedness directly from the primes. Instead, what these results suggest is that the primes activated different mechanistic theories about memory, which subsequently influenced the information that was retrieved when participants read the vignette. For example, a participant who was primed with the theory of *associations being beneficial for memory* may have judged that that the items being from the same category would create associations more readily than for unrelated items, thus becoming more likely to predict a relatedness benefit. Conversely, a participant who was primed with the theory of *distinctiveness being beneficial for memory* may have judged that the related items would be less distinctive than the unrelated items, thus becoming more inclined to predict a relatedness cost. Remarkably, the distinctiveness prime was the only manipulation that was successful at finally eliminating the predicted relatedness benefit in location memory. That said, it was not enough to produce a cost overall.

Although the priming manipulation in Experiment 8 does not entirely rule out the possibility that participants' belief explanations could be post-hoc justifications, it does clearly demonstrate that the prior activation of different mechanistic memory theories can lead to different relatedness beliefs being expressed. Furthermore, comparing how participants' free-response reasoning responses were classified to how they self-classified their responses (Table 19) suggests that even when they did base their predictions on particular theories, many of them still had difficulty clearly articulating the ideas. Thus, the large number of participants expressing an unexplained related/unrelated preference in the earlier experiments should not be taken to mean that these ideas were *not* activated.

General Discussion

The goal of the present experiments was to investigate participants' metamemory beliefs regarding the effects of semantic relatedness. In all experiments, I found that participants tended to express the belief that a memory task involving a related list of words would be easier to remember than one involving an unrelated list of words. This effect was obtained in a task where this is indeed the case (item memory) but also in a task where the opposite is the case (location memory). Thus, overall participants appeared to exhibit a kind of "relatedness halo" in their predictions, predicting that relatedness is beneficial to memory even in a context where it has been demonstrated to be harmful (Lu et al., 2023). Indeed, across all experiments, eliminating this belief in a benefit of relatedness was difficult and I never found a condition that led to a significant belief in a relatedness cost.

The relatedness halo across task and elicitation context

Overall, metamemory beliefs about the effect of relatedness tended to be positive but were also to some extent task-dependent. In Experiments 4 and 7, when participants were asked to make their predictions for either an item memory task or a location memory, their predicted relatedness benefit was greater overall for item than for location. Looking beyond the aggregate, however, revealed that this effect was driven by more participants predicting a relatedness benefit for item, and more participants predicting a cost for location (the proportion of participants predicting no difference did not differ). Within each belief direction subset (benefit or cost), the magnitude of the predicted effect was the same across task. That is, individuals were *not* predicting the same effect of relatedness for both tasks except to a lesser extent in the location task. Instead, the task descriptions influenced how likely it was that they would end up predicting that relatedness would result in a benefit, a cost, or no effect.

Experiment 6 provided further evidence that individuals were not sensitive to the moderating effect of task on relatedness. When asked to make relatedness predictions for both the item task and the location task, a context in which task should have been the most salient, the aggregate prediction was a similar-sized benefit of relatedness for both tasks, with no interaction. Participants' relatedness predictions were highly concordant across the item and location tasks: A majority predicted a relatedness effect in the same direction for both tasks, and their relatedness predictions were also significantly correlated across task. Within the framework, the two task descriptions when presented together appeared to activate the same naïve theory, which tended to lead to either a benefit of relatedness or no difference.

What leads to the expression of different relatedness predictions?

Experiments 4, 7, and 8 demonstrated that people tend to invoke different explanations for their relatedness predictions (for a benefit vs. a cost vs. no difference). People who predicted a relatedness benefit tended to reason in terms of associations, connections, and similarity-based cueing—that it would be easier to mentally connect items within the same category. On the other hand, people who predicted a relatedness cost tended to reason in terms of distinctiveness or interference—that items within the same category would more easily blur together or be confused with each other, and items in different categories would more readily stand out. Finally, people who predicted no effect of relatedness tended to explain this in terms of their own memory abilities being the limiting factor claiming, for example, that their memory is so poor that item relatedness would not matter. Experiment 8 provided evidence that participants' predictions were the result of these different lines of thought: When I directly primed participants with one of two different ideas about how memory operates, this changed the proportion of participants expressing different kinds of beliefs. Inducing participants to think about associations in memory led more people to predict a relatedness benefit in the location memory task; inducing them to think about distinctiveness in memory led more people to predict a relatedness cost.

I have proposed a framework that conceptualizes each task vignette as a set of memory cues that lead to the activation/retrieval of different stored information, and consequently lead to different judgments. Within this framework, the differences that we observe across each task description (i.e., item vs. location vs. both) arose because each one consists of a slightly different set of cues, such that participants are more likely to traverse down one cognitive pathway or another. Some participants, for example, may note that the location task demands distinguishing between similar items and therefore use distinctiveness as a criterion, and end up predicting a cost. Other participants have a relatively inflexible view of their memory and use their own memory abilities as the main criterion. Once on a particular path, the anticipated magnitude of the benefit or cost was more or less the same between tasks. That is, individuals were *not* predicting the same effect of relatedness just to a lesser extent in the location task. Instead, the cues present in each task description influenced the proportion of participants following a particular cognitive path.

Given that the majority of participants predicted a benefit of relatedness, even for the location memory task, it appears evident that the “benefit” pathway is a stronger attractor than the “cost” pathway. When cued with the concept of semantic relatedness in a memory task, participants’ default narratives tend to align more closely with association-based ideas (that lead to an expected benefit) rather than with distinctiveness-based ideas (that lead to an expected cost). The observation that participants who predicted a benefit were the most confident about it further suggests that the “benefit” path may be associated with more familiar or more easily accessible ideas. That is, when cued with the task descriptions, concepts such as *associations/connections* are more readily retrieved than concepts such as *distinctiveness/interference*. This idea also fits with the observation that not even the distinctiveness prime was sufficient to shift participants to a relatedness cost in the aggregate. Thus, when faced with a memory task, many participants invoke a particular naïve theory—relatedness facilitates associations, hence it benefits memory. Why does this particular theory tend to be the “default”? One can only speculate, but associative learning principles likely are more common in everyday memory experiences, whereas the concept of distinctiveness may be comparatively rarer.

Metamemory beliefs as constructive, context-dependent judgments

The idea emerging from the present work is that the metacognitive "beliefs" reported via memory belief questionnaires, as used by us here as well as by others in the literature (e.g., Koriat et al., 2004; Mueller et al., 2013), are probably best considered to be "constructions" rather than fixed entities that are being retrieved. A similar distinction has been made in research on preferences. Slovic (1995) contrasted the constructivist notion of preferences with the layman's idea of preferences as fixed entities to be reported, as follows: "… I can describe three different views regarding the nature of values. First, values exist—like body temperature—and people perceive and report them as best they can, possibly with bias […] Second, people know their values and preferences directly—as they know the multiplication table […] Third, values or preferences are commonly constructed in the process of elicitation". This is not a completely novel claim. Cavanaugh et al. (1998) has previously argued that responses to memory belief questions are constructed from information that can be retrieved from memory at the time of judgment. Which information is available, accessible, and subsequently used to construct a response is heavily influenced by the context of the question and by individual differences such as in motivation.

Here I provide evidence for this idea in the context of metamemory beliefs about how factors such as relatedness influence memory. The item and location task descriptions influenced the likelihood of participants predicting a benefit, a cost, or no effect of relatedness. Priming participants with different mechanistic theories about memory significantly altered the likelihood that different relatedness beliefs would be expressed. These results suggest that participants' relatedness beliefs were not pre-formed but were instead actively constructed in response to the information presented during the experiment.

Previously, Koriat et al. (2004) has lamented that “understanding fully how and when theory-based knowledge is accessed and combined with experience-based subjective knowledge may require nothing less than a theory of how beliefs are organized and activated”. Here I have outlined a framework that conceptualizes metamemory beliefs as cue-dependent judgments, which I hope will productively guide future metacognition research. For example, one avenue for potential investigation concerns the dissociation between expressed beliefs and item-by-item judgments of learning (e.g., Koriat et al., 2004; Kornell et al., 2011). I hypothesize that this divergence may be accounted for by the differences in elicitation context in which these two kinds of judgments are constructed and expressed. I emphasize that metacognition research must consider the importance of elicitation context for judgments and the constructive nature of those judgments.

Conclusion

The present investigation provides insight into the nature of individuals’ beliefs about how relatedness influences memory, and why metamemory beliefs can be misaligned with objective performance. I propose that metamemory beliefs should be thought of as flexible and context-dependent judgments that are constructed online at the time of elicitation from available and accessible information.

Chapter 3

As Chapter 2 has established, most people believe that related items are easier to remember than unrelated items, even in the context of a location memory task where relatedness has an objective memory cost. In Chapter 3, across three pre-registered experiments, I show that this predilection diminished after first-hand experience with the location memory task: Participants' predictions showed a decrease in the predicted benefit of relatedness, although they did not reverse to predicting a cost of relatedness in the aggregate. Critically, changes in their belief reports followed the same direction as their task experience: Those who experienced a benefit shifted their relatedness predictions upward accordingly whereas those who experienced a cost of relatedness shifted theirs downward. Across all experiments, participants' final predictions about the effects of relatedness on memory were predicted by their objective relatedness experiences in the task. In Experiments 9 and 10, I found that previous predictions continued to influence their final predictions after accounting for the effect of task experience, although this effect was not significant in Experiment 11. These results are accounted for by a framework that construes participants' elicited memory beliefs as judgments that are constructed and updated based on the cues and information that are available in the elicitation context.

Metacognition refers to the ability to monitor and control our cognitive processes, i.e., thinking about our own thinking. Metacognitive acts of control abound in everyday life, such as deciding whether to study more for a test, or to create a reminder for an appointment. Theoretical accounts of metacognition, such as the dual-basis view (Koriat, 1997; Koriat et al., 2004) and analytic processing theory (Mueller & Dunlosky, 2017), emphasize a central role for *a priori* theories or beliefs about memory in our metacognitive judgments. As one illustration, people tend to believe that related items are easier to remember than unrelated items, which in part leads them to predict these items that related items will be better learned (Mueller et al., 2013).

However, metamemory beliefs are sometimes misaligned with objective performance, which consequently leads to suboptimal control decisions. For example, people often erroneously believe that blocked study schedules (e.g., AAABBBCCC) will be more effective for learning categories than interleaved schedules (e.g., ABCACBCBA; McCabe, 2011; Yan et al., 2016), and subsequently they prefer to learn in a blocked fashion even when this schedule is not ideal (e.g., Lu et al., 2021; Tauber et al., 2013). People also tend to believe that related items are easier to learn in a location memory task and consequently spend less time studying these items, even when they perform worse in this condition (Lu et al., 2023; Lu & Risko, under review). In both of these cases, learners' metamemory beliefs run contrary to their objective memory performance, leading to counterproductive behavior. Previous research has shown that these erroneous beliefs can be corrected with subsequent experience during the experiment (Yan et al., 2016; Koriat & Bjork, 2006). In this chapter, I investigate how and why participants' metamemory beliefs are updated or changed over the course of an experiment.

The anatomy of a metamemory belief

Although many theoretical accounts of metacognition assume a central role for metamemory beliefs or theory-based reasoning, comparatively little work has been done to interrogate the nature of these beliefs. In the dual-basis view (Koriat, 1997; Koriat et al., 2004), participants can rely on *a priori* theories (i.e., beliefs about memory) and/or experience-based cues (e.g., processing fluency/ease, subjective feelings of knowing) when asked to make judgments of their learning. Within analytic processing theory (Mueller & Dunlosky, 2017), a belief is some proposition or representation that either exists prior to the learning context or is formed online through experiences during the task (Dunlosky et al., 2015). Dunlosky and colleagues cited the belief about relatedness (that related items are easier to remember) as an example of a belief that participants have formed based on what they have been taught in school or through everyday experiences, while beliefs about encoding strategies (the efficacy of imagery vs. rote repetition; Hertzog et al., 2008) may exemplify a belief that is formed during the course of the experiment.

In the metacognitive literature, no one really denies that we sometimes rely on *a priori* beliefs when making judgments of learning (though beliefs are not always automatically activated: Koriat et al., 2004, Kornell et al., 2011). Most researchers have measured metamemory beliefs using a questionnaire method, usually by presenting people with a description of a hypothetical memory experiment and asking them to predict their performance in the task under certain conditions (e.g., Koriat et al., 2004, Kornell et al., 2011; Lu & Risko, under review; Mueller et al., 2013; Witherby & Tauber, 2017). Another approach involves asking people to make pre-study judgments of learning, wherein participants are told only that an upcoming to-be-learned item has a certain characteristic (e.g., the word pair you are about to study is related; Price & Harrison, 2017). By pre-empting any experience associated with processing the item, pre-study judgments are assumed to be a pure reflection of one's beliefs about the characteristic in question. Finally, yet another method involves some variation of a learner-observer task, where some participants observe another learner's study trials and then make judgments of that person's learning. Critically, the observation group views all to-be-learned stimuli in the same format (e.g., the same study duration or font size) as the learner, but with the actual stimuli replaced by meaningless letter strings (e.g., Yang et al. 2018) or a black rectangle (e.g., Mueller et al. 2014). As with pre-study judgments of learning, since the observers do not view the actual stimuli, their judgments are assumed to reflect only their beliefs about, say, font size. In one creative variant of this task, unbeknownst to the participant, the other person's trials that they are observing are in fact their own trials from an earlier phase of the experiment (Yang et al., 2018).

Yang and colleagues argue that it is insufficient to show that beliefs directionally agree with judgments of learning; instead, we should attempt to quantify the contribution of beliefs through mediation and regression analyses (Yang et al., 2021). Although they identify some methodological issues with each of the aforementioned measures of belief, the assumption (measurement error aside) is still that they adequately tap into pre-existing beliefs about memory. However, the overarching challenge remains that we lack a unified theory of beliefs and how they are organized and activated (Koriat et al., 2004). Attempts to measure the contribution of beliefs to metamemory judgments will remain elusive because the relation between the beliefs and judgments of learning has been shown to vary dramatically. For example, participants' pre-study judgments of learning have been shown to correlate nearly perfectly ($r > .90$) with their immediate (post-item presentation) judgments of learning (e.g., Price & Harrison, 2017). On the other hand, sometimes participants' questionnaire responses indicate a belief that a particular factor benefits memory, but this is not expressed in item-by-item judgments (e.g., Kornell et al., 2011).

Belief reports as constructive, cue-dependent judgments

In light of the aforementioned issues, Lu and Risko (under review) have recently proposed a framework that conceptualizes metamemory belief reports as judgments that are constructed in response to a specific elicitation context. What kind of information is available, accessible, and subsequently used to construct a response is heavily influenced both by the context of the question and by individual differences (see also Cavanaugh et al., 1998). Lu and Risko argued that when a belief report is solicited in a questionnaire, the task vignette and prompt form a set of memory cues that lead to the activation/retrieval of stored information which forms the basis of the judgment.

According to this framework, any measure of belief will be context- and cue-dependent in a way that is not fundamentally different from how item-by-item judgments of learning are made (i.e., cue-utilization framework; Koriat, 1997). Where belief reports and immediate judgments of learning differ is in the kinds of information that is available to them, such that participants making (pre-experiment) belief judgments have access only to prior experiences, whereas in a judgment of learning they also have access to the immediate experience of processing the item. The beliefs as constructive judgments framework provides a productive lens through which to view the role of beliefs in metacognition. For example, the finding that pre-study judgments of learning almost perfectly predict immediate post-item judgments (Price & Harrison, 2017) can be accounted for if the former judgment is still highly accessible when judgment is solicited at a later time.

The framework positions “beliefs” (or, more accurately, belief reports) about memory as judgments that are constructed on the basis of various pieces of retrieved information (Lu & Risko, under review). The constructive view of beliefs has a long tradition in social cognition research, where a “belief” is assumed to be the estimate of the likelihood that some knowledge is correct or that some event or state of affairs will occur. Our beliefs are assumed to be computed online from “belief-relevant information” that is stored in knowledge structures known as schema (Crocker et al., 1984; Schwarz & Bohner, 2001; Wyer & Albarracín, 2005). These ideas are echoed in what Mandelbaum and colleagues have recently termed “the fragmentation of belief” (Bendaña & Mandelbaum, 2021; Porot & Mandelbaum, 2021). In the fragmented model of belief storage (Bendaña & Mandelbaum, 2021), it is argued that we do not just store a single belief about *P* but multiple fragments of belief that can be inconsistent with each other (i.e., both *P* and not *P*).

Such a structure can account for various empirical phenomena from seemingly disparate areas of research. For example, people can be famously inconsistent with what they profess to believe at various times. They can report beliefs that are internally inconsistent because the information that they retrieve and is active at any given moment can be different (i.e., context-dependent). Another example is recovery from extinction in Pavlovian conditioning, such as when a mouse learns to associate a specific sound with an electric shock. The fear response to the sound can be ‘extinguished’ after it has been repeatedly paired with no shock, in the sense that the organism learns a new association between sound and no shock, which reduces the response (Pearce & Hall, 1980). The original association between the sound and shock is clearly not lost, given that reinstating the initial conditioning context cues the recovery of the fear response.

Finally, consider the “internal wisdom of the crowd” effect (Vul & Pashler, 2008; Herzog & Hertwig, 2009), the finding that averaging multiple answers from the same person reduces bias and improves accuracy compared to a single answer. Different spatiotemporal contexts cue participants to access different pieces of information; answers produced by drawing on one piece of information are likely to have different sources of error than those produced by drawing on another piece of information. This also accounts for why people who have a larger delay between their first answer and their second are more accurate (Vul & Pashler, 2008): The longer delay reduces reliance on the first answer and allows for other information to be activated, which typically increases the independence of the sources of error.

Mandelbaum and colleagues posit a propositional structure for beliefs, such that belief fragments are essentially propositionally structured strings of mental representations (Porot & Mandelbaum, 2019; Quilty-Dunn & Mandelbaum, 2018). A philosophical treatment of the nature of beliefs is beyond the scope of the current article, but our framework is compatible with the notion that some knowledge is stored in a propositional format in the mind, and that this knowledge can form the basis of belief responses. However, I argue that belief reports can be constructed on the basis of other kinds of retrieved and available information, of which the contents can be propositional (e.g., previous judgments, semantic information) or non-propositional (e.g., specific episodes, perceptual experiences). In the current investigation, I examine whether the framework can capture how metamemory belief reports change over the course of an experiment as participants acquire new information from the experience.

Previous investigations of metamemory belief change

I begin by reviewing two previous studies that have examined how metamemory biases (i.e., a mismatch between participants’ predictions and their performance) can be remedied with experience. Koriat and Bjork (2006a; 2006b) examined the *foresight bias*: the tendency for people to fail to predict how difficult an answer will be to generate on a later test, because the answer is in front of them when they are making a judgment of learning (Koriat & Bjork, 2005). For example, when learning paired associates for a later cued-recall test, participants give similar judgments of learning to the pairs *cheese-cheddar* and *cheddar-cheese*, even though the latter (*cheddar*-____; forward associated pair) is far easier to recall than the former (*cheese*-____; backward associated pair). Koriat and Bjork (2006a) reported that the bias in participants' metacognitive judgments was reduced with repeated test experience, but not with study experience alone. They argued that test experience was effective at debiasing participants because they were able to learn which cue would help memory during the test and which would not (i.e., forward association strength vs. backwards association strength). That is, participants gained information about the performance *diagnosticity* of each cue, which improved judgment accuracy because they were able to discount the fluency from the misleading cue.

Yan and colleagues (2016) examined participants' tendency to believe that blocked schedules (e.g., AAABBBCCC) are more effective for learning categories than interleaved schedules (e.g., ABCACBCBA). Previous studies have firmly established that participants tend to express this belief, which often runs contrary to their actual learning performance being better with interleaving (Kornell & Bjork, 2008; Lu et al., 2021; McCabe, 2011; Tauber et al., 2013). Yan and colleagues (2016) reported a series of interventions that attempted to overturn this erroneous belief in the superiority of blocking, which proved to be remarkably difficult. These interventions included giving participants study and test experience, as well as explicitly telling them that interleaving improved the ability to detect differences between the categories. Only in their final experiment, where they separated test performance on the two schedules by presenting them in separate study-test blocks (Experiment 6), did this lead to a majority of learners indicating a preference for interleaving. Note, however, than Yan and colleagues (2016) did not solicit pre-experience belief reports. Therefore, while their interventions clearly had some influence on post-task judgments, they could not examine how beliefs changed, nor to what degree prior beliefs might be combined with new information. However, the assumption is that participants' schedule preference was only altered when they could unambiguously attribute superior performance to the interleaved schedule, thus gaining information about the relative efficiency of the two schedules.

The current investigation

In the present investigation, I examined how participants' metamemory belief reports changed with experience through the lens of the beliefs as constructive judgments framework (Lu & Risko, under review). In this framework, belief reports are constructed from information that can be retrieved from memory at the time of judgment. Thus, participants' experience of the memory test is information that is expected to be weighted heavily in this constructive process because it is not only recent but also highly relevant (i.e., diagnostic) and highly accessible.

I report three pre-registered experiments (Experiment 9: https://osf.io/kjwxt/; Experiment 10: https://osf.io/x7ra6; Experiment 11: https://osf.io/xnvep) that solicited participants' predictions of how item relatedness would influence their performance in a memory task, both before and after experiencing the memory task. All three experiments followed a similar procedure. Participants first were presented with a vignette that described a hypothetical memory experiment where to-be-remembered words would be presented within grid displays (e.g., Lu et al., 2023; Lu et al., 2024; Lu & Risko, under review). In one condition, the grid display would contain words that were all from the same category (Related); in the other condition, the grid display would contain words that were all from different assorted categories (Unrelated). After reading the vignette and providing their predictions for the related and unrelated conditions, participants then went through the actual memory task that had been described to them. Similar to Experiment 6 in Yan et al. (2016), the location memory task has separate study-test blocks for the related and unrelated conditions, so we should expect little ambiguity or confusion in terms of performance in each condition. After experiencing both conditions of the memory task for themselves, participants were again asked to provide their predictions for the related and unrelated conditions.

In Experiment 9, participants were told either that the memory test would be a location memory test (Location Memory group) or that it would be an item recall test (Item Memory group). Following their predictions, they were then tasked with performing the respective memory task (i.e., item vs. location task was manipulated between-subjects). While the related condition has been found to be associated with a benefit for the item memory task, it results in a cost for the location memory task (Lu et al., 2023). Despite this objective result, participants tend to predict the opposite in the location memory task – that relatedness will be beneficial to memory (Lu et al., 2023; Lu & Risko, under review). I expected to replicate the dissociation between participants' pre-task beliefs and their objective memory performance. That is, I anticipated that participants would exhibit a relatedness benefit for item memory and a relatedness cost for location memory, but express a relatedness benefit for both tasks in their pre-task predictions (though to a lesser extent for location than item memory; Lu et al., 2023; Lu & Risko, under review).

The novel hypotheses of interest were concerned with participants' post-task predictions: I predicted that their experience of the memory test would be weighted heavily in these predictions. First, I hypothesized that participants' post-task predictions would be changed or updated from their pre-task predictions after experience with the memory task. I predicted that there would be a reduction in the predicted relatedness benefit for participants experiencing the location memory test, but not for those experiencing the item memory test. Second, I hypothesized that an individual's post-task predictions would be driven by their individual experiences in the memory test, such that participants who experienced an objective cost would be more likely to reduce their predictions of a relatedness benefit. Third, I explored to what extent participants' post-task predictions were a combined function of their pre-task predictions and their objective memory experiences. While I expected objective memory performance to be a significant influence on these judgments, I wondered to what extent previously retrieved information or judgments would be a contributing factor. Finally, I explored participants' self-reported reasoning for their post-task predictions, expecting that they would cite their experiences in the memory test as the basis of these judgments.

Experiment 9

In Experiment 9, participants provided related and unrelated predictions at two time points: pre-task and post-task. Participants were also asked to provide estimates of their performance in the task that they had experienced (postdictions).

Method

Participants. Data from 240 participants (135 women, 102 men, 1 other, 2 unknown, *M* = 38.51 years, *SD* = 12.36) were analyzed. A power calculation performed with Superpower's *ANOVA_exact* (Lakens & Caldwell, 2021) estimated that this sample size would be sufficient to achieve 80% power to detect a task by relatedness interaction in participants' pre-task predictions[7]. Participants were recruited from Prolific and compensated 1.90 GBP for 15 minutes. Participants were randomly assigned to either the item memory group (*N* = 122) or the location memory group (*N* = 118).

Materials. Stimuli were drawn from eleven categories of household object words (ten words per category): wearables, tools, toiletries, office supplies, kitchen supplies, musical instruments, toys, sports equipment, furniture, pantry items, and cleaning supplies. For each participant, the related items display was populated with ten items from a single, randomly chosen category, and the unrelated items display was populated by randomly selecting one item from each of the remaining ten categories. Examples of the two types of grid displays are shown in Figure 1 (note that participants never saw a grid with all items visible at once). As each 5 x 5 grid had 25 clickable location squares, 10 of these contained items and 15 were empty squares. Item positions in each grid were pseudo-randomly assigned such that each row and column contained two items (cf. Siegel & Castel, 2018).

Procedure. Each participant was provided with a brief experimental vignette that outlined the memory task, followed by a pre-task probe that asked them to predict item memory performance (*N* = 122) or location memory performance (*N* = 118). This was followed by the location memory task proper, the postdiction probe, and finally the post-task prediction probe. Figure 2 shows a schematic diagram of the procedure. These tasks are described below.

[7] I entered the following estimates: M_Item_Related = 5.72, M_Location_Related = 5.20, M_Item_Unrelated = 4.72, M_Location_Unrelated = 4.81, within-subjects *r* = 0.45, pooled *SD* = 1.69. These estimates were obtained from previous experiments.

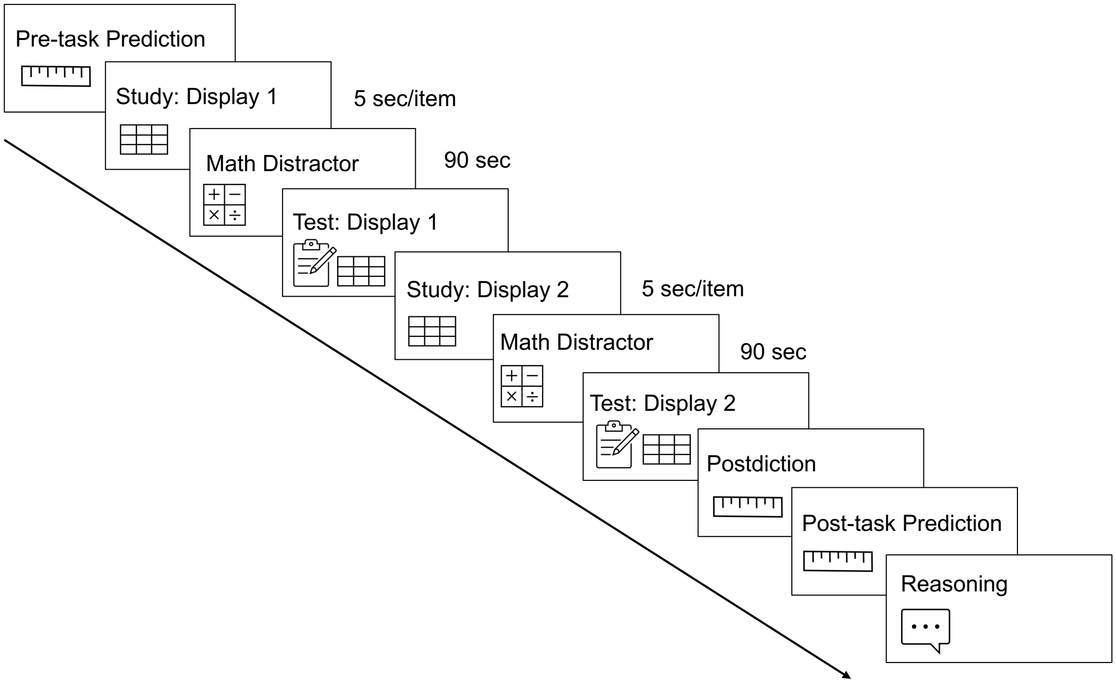


Figure 9. Schematic diagram of Experiment 9 procedure.

Experimental Vignette. Participants were given the following description, along with some example pictures of the encoding and test tasks:

> "In this task, you will be presented with 10 words in a grid display (see below). In the learning phase, each word will appear on the grid one at a time. Please try to remember [each word/the location of each word] that was presented. Once you have seen all of the words, your memory will be tested. In the memory test, you will be given [a blank text box and asked to recall all the words that were presented to you/each target word one at a time and you will have to indicate its location in the grid]."

Pre-task Prediction. Participants were given the following instruction:

> "Please read these instructions carefully. As noted previously, you will complete a task that involves learning [words/the locations of words] presented in a grid display. Some

displays contain words that are all in the same category (e.g., all words are kinds of FURNITURE) and some displays contain words from different assorted categories (e.g., one word might be a kind of FURNITURE, one word might be a TOOL, another word might be a TOY). For each type of display, on the scales below please estimate the number of words (out of 10) [which you will be able to correctly recall/for which you will be able to correctly recall the location]."

On the same page, participants made their ratings on two Likert scales: "all items from same category" and "all items from different categories".

Encoding Task. On each trial, a single item in the grid was shown to the participant (Figure 7A). The visible item was presented for 5000 ms in white text on a black square. A text prompt on the right reminded participants to remember the item (item group) or the item's location (location group). There were 10 trials for a single grid (one for each item, presented in random order) and the intertrial interval was 400 ms.

Test Task (Item Recall or Location Memory). Participants in the item memory group were asked to type all the words that they could remember from the display into an on-screen text field (Figure 7B). For participants in the location memory group, on each trial they were shown an empty grid on the left with a target item shown on the right (Figure 7C). They were instructed to click on the square that corresponded to the original location of that target item. There were 10 trials for a single grid (one for each item, presented in random order) and no feedback was given.

Performance Postdiction. Participants were given the following instruction:

"Please read these instructions carefully. As noted previously, in the memory test you were given [a blank text box and asked to recall all the words that were presented to you/each target word one at a time and you had to indicate its location in the grid]. For

> each type of display, on the scales below please estimate the number of words (out of 10) [which you were able to correctly recall/for which you were able to correctly recall the locations]."

The Likert scales were identical to the first pre-task prediction probe.

Comprehension Check #1. Participants were asked to respond to a multiple-choice question: "On the previous page, you were to make estimates about: (i) your performance in an upcoming task, (ii) your performance in a previous task, or (ii) it did not specify."

Post-task Prediction and Comprehension Check #2. Participants were told that they would be learning a new display, and to make predictions for each of the two types of display. Instructions and Likert scales were identical to the first pre-task prediction probe. On the next page, participants were given a comprehension check question identical to the first.

Self-Reported Reasoning. Participants were given the following instruction and asked to type their responses into an on-screen text field:

> "In the last task we asked you to make predictions about the number of words for which you would be able to correctly [recall/recall the location]. You estimated [X] when all items are from the same category, and you estimated [Y] when all items are from different categories. Please tell us about your reason(s) for selecting the number of items for each display type. There is no right or wrong answer, as we're interested in how individuals make predictions about memory."

Finally, participants were asked to self-declare if they were paying attention during the task and asked to provide their age and gender.

Results

Data from 60 participants were not analyzed according to the exclusion criteria set in the pre-registration: (1) self-reported that that they were not paying attention or did not give effort during the task; (2) did not answer at least 14 math questions and achieve at least 70% correct in each math distractor task; and/or (3) did not answer the final two comprehension checks correctly. An additional 4 participants who took part after the stopping rule were also excluded. The final sample size was $N = 240$. Data and analysis code for all experiments are available at https://osf.io/4ary5/files.

Objective Memory Performance

I conducted a 2 (Relatedness: Related vs. Unrelated Items; within) x 2 (Memory Task: Item vs. Location; between) mixed ANOVA on participants' objective memory performance. The analysis revealed a main effect of memory task, $F(1,238) = 62.72$, $MSE = 0.07$, $p < .001$, $\eta_G^2 = .160$, such that item memory performance was better than location memory performance. There also was a main effect of relatedness, $F(1,238) = 8.51$, $MSE = 0.03$, $p = .004$, $\eta_G^2 = .010$, such that performance was better for the related items display. Most importantly, the interaction was significant, $F(1,238) = 65.27$, $MSE = 0.03$, $p < .001$, $\eta_G^2 = .071$. Follow-up analyses revealed that whereas related items led to better item recall performance than unrelated items, $F(1,121) = 67.42$, $MSE = 0.02$, $p < .001$, $\eta_G^2 = .147$, related items led to poorer location memory performance than unrelated items, $F(1,117) = 12.01$, $MSE = 0.03$, $p < .001$, $\eta_G^2 = .026$. Figure 4 shows a plot of objective memory performance (proportion correct) as a function of memory type and item relatedness. Table 20 shows objective memory performance (converted to the number of correct items) along with their subjective estimates of that performance (postdictions) for comparison.

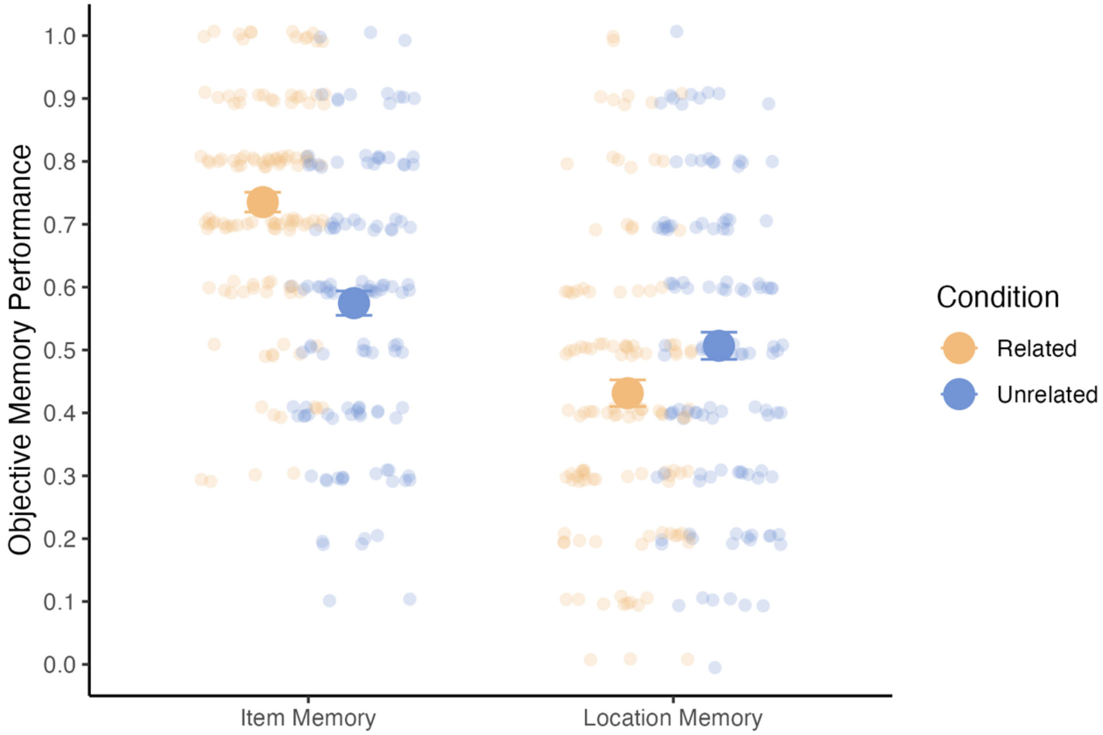


Figure 10. Experiment 9: Mean proportion of correct items recalled (item memory) and correct locations selected (location memory) in the related and unrelated conditions. Circles in the background represent individual participants' memory performance; error bars are standard errors.

Subjective Memory Predictions

I conducted the following pre-registered analyses: (1) with respect to pre-task predictions, I expected to replicate the finding that participants would predict a relatedness benefit for both item and location memory, though to a lesser extent in the latter; (2) with respect to participant's predictions changing from pre-task to post-task, I hypothesized that the relatedness benefit might decrease after experience for the location memory group. Table 21 shows participants' subjective memory predictions (pre-task and post-task).

Table 20. Experiment 9: Participants' mean (SD) memory performance and postdictions

	Item Memory (*N*=122)		**Location Memory (*N*=118)**	
	Related	Unrelated	Related	Unrelated
Performance	7.35 (1.73)	5.75 (2.14)	4.31 (2.30)	5.07 (2.35)
Postdiction	6.90 (1.83)	5.70 (1.83)	3.69 (2.11)	3.89 (2.19)

Table 21. Experiment 9: Participants' mean (SD) metamemory predictions

	Item Memory (*N*=122)		**Location Memory (*N*=118)**	
	Related	Unrelated	Related	Unrelated
Pre-task Prediction	5.90 (1.61)	4.54 (1.40)	4.88 (1.80)	4.49 (1.66)
Post-task Prediction	6.51 (1.54)	5.28 (1.61)	3.78 (1.94)	3.92 (2.10)

Pre-task Predictions. I conducted a 2 (Relatedness: Related vs. Unrelated) x 2 (Memory Type: Item vs. Location) mixed ANOVA on participants' pre-task predictions (i.e., before they had any task experience). This analysis revealed a main effect of relatedness, $F(1,238) = 66.87$, $MSE = 1.37$, $p < .001$, $\eta_G^2 = .068$, such that pre-task predictions were higher for related items than unrelated items. The main effect of memory type was also significant, $F(1,238) = 8.84$, $MSE = 3.88$, $p = .003$, $\eta_G^2 = .027$, such that pre-task predictions were higher for item memory than for location memory. The interaction was significant as well, $F(1,238) = 20.57$, $MSE = 1.37$, $p < .001$, $\eta_G^2 = .022$. Follow-up analyses revealed that item memory predictions were significantly higher for related items than for unrelated items, $F(1,238) = 82.17$, $p < .001$, $\eta_G^2 = .171$; location memory predictions were also significantly higher for related items than for unrelated items, $F(1,238) = 6.52$, $p = .011$, $\eta_G^2 = .013$, although this effect was smaller than that for item memory predictions.

Change from Pre-task to Post-task Predictions[8]. To examine whether participants changed their beliefs after task experience, I conducted a 2 (Judgment Time: Pre-task vs. Post-task) x 2 (Relatedness: Related vs. Unrelated) x 2 (Memory Type: Item vs. Location) mixed-subjects ANOVA on participant predictions. This analysis revealed a main effect of relatedness, $F(1,238) = 89.46$, $MSE = 1.35$, $p < .001$, $\eta_G^2 = .041$, such that participants' predictions were higher for related items than for unrelated items. There was also a main effect of memory type, $F(1,238) = 60.93$, $MSE = 6.54$, $p < .001$, $\eta_G^2 = .124$, such that participants' predictions were higher for item memory than for location memory. There was no main effect of judgment time, $F(1,238) = 0.59$, $MSE = 2.67$, $p = .441$, $\eta_G^2 < .001$. While the three-way interaction was not significant, $F(1,238) = 1.97$, $MSE = 1.23$, $p = .161$, $\eta_G^2 < .01$, I follow up with separate two-way ANOVAs for the item memory and location memory groups below for ease of exposition and for the *a priori* reason that the two groups experienced different memory tasks.

In the item memory group, the two-way interaction between relatedness and judgment time was not significant, $F(1,121) = 0.47$, $MSE = 1.12$, $p = .494$, $\eta_G^2 < .001$. Related items were predicted to be better remembered than unrelated items overall, $F(1,121) = 165.20$, $MSE = 1.24$, $p < .001$, $\eta_G^2 = .151$, and post-task predictions were higher than pre-task predictions overall, $F(1,121) = 23.05$, $MSE = 2.39$, $p < .001$, $\eta_G^2 = .046$. In the location memory group, there was no main effect of relatedness, $F(1,117) = 1.22$, $MSE = 1.46$, $p = .272$, $\eta_G^2 = .001$, and post-task predictions were lower than pre-task predictions overall, $F(1,117) = 27.84$, $MSE = 2.95$, $p < .001$, $\eta_G^2 = .047$. Critically, the two-way interaction between relatedness and judgment time was significant, $F(1,117) = 6.21$, $MSE = 1.35$, $p = .014$, $\eta_G^2 = .005$. At the pre-task timepoint, related items were predicted to be better remembered than unrelated items, $F(1,117) = 5.79$, $MSE =$

[8] While I also pre-registered an analysis of the change from pre-task to postdictions, I do not report this as it was not meaningfully different from the change from pre-task to post-task predictions.

1.55, p = .018, η_G^2 = .013; at the post-task timepoints, this difference was no longer significant, F(1,117) = 0.97, MSE = 1.27, p = .328, η_G^2 = .001.

Relatedness Belief Direction. The above analyses revealed a smaller predicted benefit of relatedness for the location memory task (vs. the item memory task) at the pre-task timepoint. However, the aggregate results could mask important differences in participants' beliefs and the distribution of those beliefs.

Lu and Risko (under review) found that the task by relatedness interaction (before task experience) was driven by more participants predicting a benefit for item, and more participants predicting a cost for location (i.e., a difference in belief kind rather than degree). Thus, based on participants' related vs. unrelated prediction differences, I conducted exploratory analyses after sorting them into three groups: (1) those who predicted that they would remember more items/item locations for the related items (related benefit), (2) those who predicted that they would remember fewer items/item locations for the related items (related cost), and (3) those who predicted that they would remember equal numbers for related vs. unrelated items/item locations (no difference). Table 22 shows the proportion of participants in each group at pre-task and post-task.

Exploratory chi-square tests revealed that the proportion of participants in each group differed for item vs. location memory, at both the pre-task, $\chi^2(2)$ = 33.28, p < .001, and the post-task, $\chi^2(2)$ = 38.67, p < .001, timepoints. Pairwise comparisons (Bonferroni-adjusted) suggested that at the pre-task timepoint, more participants predicted a relatedness benefit for item memory than for location memory, $\chi^2(1)$ = 28.16, p < .001, whereas more participants predicted a relatedness cost for location memory than for item memory, $\chi^2(1)$ = 14.21, p < .001, and more participants predicted no difference in relatedness for location memory than item memory, $\chi^2(1)$

= 8.66, $p = .003$. These differences were also significant at the post-task timepoint (benefit: $\chi^2(1) = 36.22$, $p < .001$, cost: $\chi^2(1) = 8.31$, $p = .004$; no difference: $\chi^2(1) = 15.07$, $p < .001$).

To quantify whether the proportion of participants expressing each type of relatedness belief changed after experience, a generalized estimating equation (Touloumis, 2015) was used to model beliefs about relatedness as an ordinal multinomial response (benefit > no difference > cost). The two predictors were prediction time (pre-task vs. post-task) and judgment type (item vs. location), and subject was the clustering variable. The interaction was not significant ($p = .838$) and hence was removed from the model. I found that the proportion of participants in each group (benefit, no difference, and cost) differed significantly for item vs. location predictions, $b = 1.56$, 95% CI [1.15, 1.96], $z = 7.54$, $p < .001$, which parallels the results of the chi-square tests. Critically, the proportion of participants in each group also changed significantly from pre-task to post-task, $b = 0.49$, 95% CI [0.16, 0.83], $z = 2.93$, $p = .003$. In other words, there was a significant tendency for participants' expressed belief direction to change after task experience.

Table 22. Experiment 9: Proportions of participants who predicted a related benefit, related cost, or no difference at the pre-task and post-task timepoints

	Item Memory (*N*=122)			**Location Memory (*N*=118)**		
	Related Benefit	**Related Cost**	**No Difference**	**Related Benefit**	**Related Cost**	**No Difference**
Pre-task Prediction	73.8%	3.4%	23.0%	39.0%	18.9%	41.5%
Postdiction	64.8%	15.3%	20.5%	32.2%	30.3%	36.4%
Post-task Prediction	63.1%	7.6%	29.5%	23.7%	20.5%	55.1%

Relation between Objective Memory and Subjective Memory

Correlations between objective and subjective memory. I calculated related – unrelated difference scores for each participant's predictions, postdictions, and memory performance. A positive score indicates a subjective (prediction/postdiction) or objective (experienced) relatedness "benefit" and a negative score indicates a subjective or objective "cost". Tables 23 and 24 show the bivariate correlations between participants' relatedness performance, pre-task prediction, post-task prediction, and postdiction.

Table 23. Experiment 9: Bivariate correlations for item memory group

	Item Memory (*N*=122)			
	Objective Performance	**Pre-task Prediction**	**Postdiction**	**Post-task Prediction**
Objective Performance	-			
Pre-task Prediction	-.22 **	-		
Postdiction	.66 ***	-.23 **	-	
Post-task Prediction	.58 ***	.05	.65 ***	-

Note. * *p* < .05, ** *p* < .01, *** *p* < .001

Table 24. Experiment 9: Bivariate correlations for location memory group

	Location Memory (*N*=117)			
	Objective Performance	**Pre-task Prediction**	**Postdiction**	**Post-task Prediction**
Objective Performance	-			
Pre-task Prediction	-.09	-		
Postdiction	.47 ***	.13	-	
Post-task Prediction	.47 ***	.18 *	.87 ***	-

Note. * *p* < .05, ** *p* < .01, *** *p* < .001

Objective Memory and Subjective Belief Change. I hypothesized that the change in participants' *subjective* memory predictions could have been driven by their *objective* memory performance. In other words, if a participant experienced a relatedness benefit in the memory task, they might be more likely to increase the predicted benefit of relatedness, but if a participant experienced a relatedness cost, they might be more likely to reduce their initial belief in a benefit of relatedness. To explore this idea, I performed a linear regression using participants' related vs. unrelated performance differences in the memory task to predict changes in their subjective memory ratings for the related vs. unrelated conditions.[9] Belief change was operationalized as the difference between the relatedness belief (related – unrelated prediction) at pre-task and at post-task (difference of differences). Regressions were first performed separately for the item and location memory groups. For both groups, there was a positive relation (item memory: b = 5.62 [95% CI: 4.18, 7.07], t = 7.70, p < .001; location memory: b = 3.66 [95% CI: 2.28, 5.04], t = 5.26, p < .001) between their experienced objective memory performance difference (related – unrelated memory performance) and their subjective belief change in the benefit of relatedness (change in related – unrelated predictions from pre-task to post-task). In other words, participants who experienced a relatedness benefit in the memory task were likely to increase their predictions in the direction of a benefit whereas participants who experienced a relatedness cost in the memory task were more likely to decrease their predictions of a benefit.

When I conducted a single linear regression on the combined item and location data, with memory type as a factor, I found that the interaction between objective performance and memory type was marginally significant, b = -1.96 [95% CI: -3.95, 0.02], t = 1.95, p = .053, suggesting that objective performance may have had a stronger effect on belief change in the item memory

[9] One extreme value in the location condition was removed after visual inspection of the regression plot revealed this outlier.

group (i.e., *b* = 5.62 for item vs. *b* = 3.66 for location). This could be attributable to the item recall participants having had more insight into their performance compared to the location memory participants (because no feedback was given). The location group was associated with more relatedness belief change, *b* = 0.79 [95% CI: 0.28, 1.29], *t* = 3.05, *p* = .003, though both groups' belief change was predicted by their objective performance, *b* = 4.57 [95% CI: 3.57, 5.57], *t* = 9.03, *p* < .001. Figure 5 illustrates the relation between objective memory performance and subjective memory change.

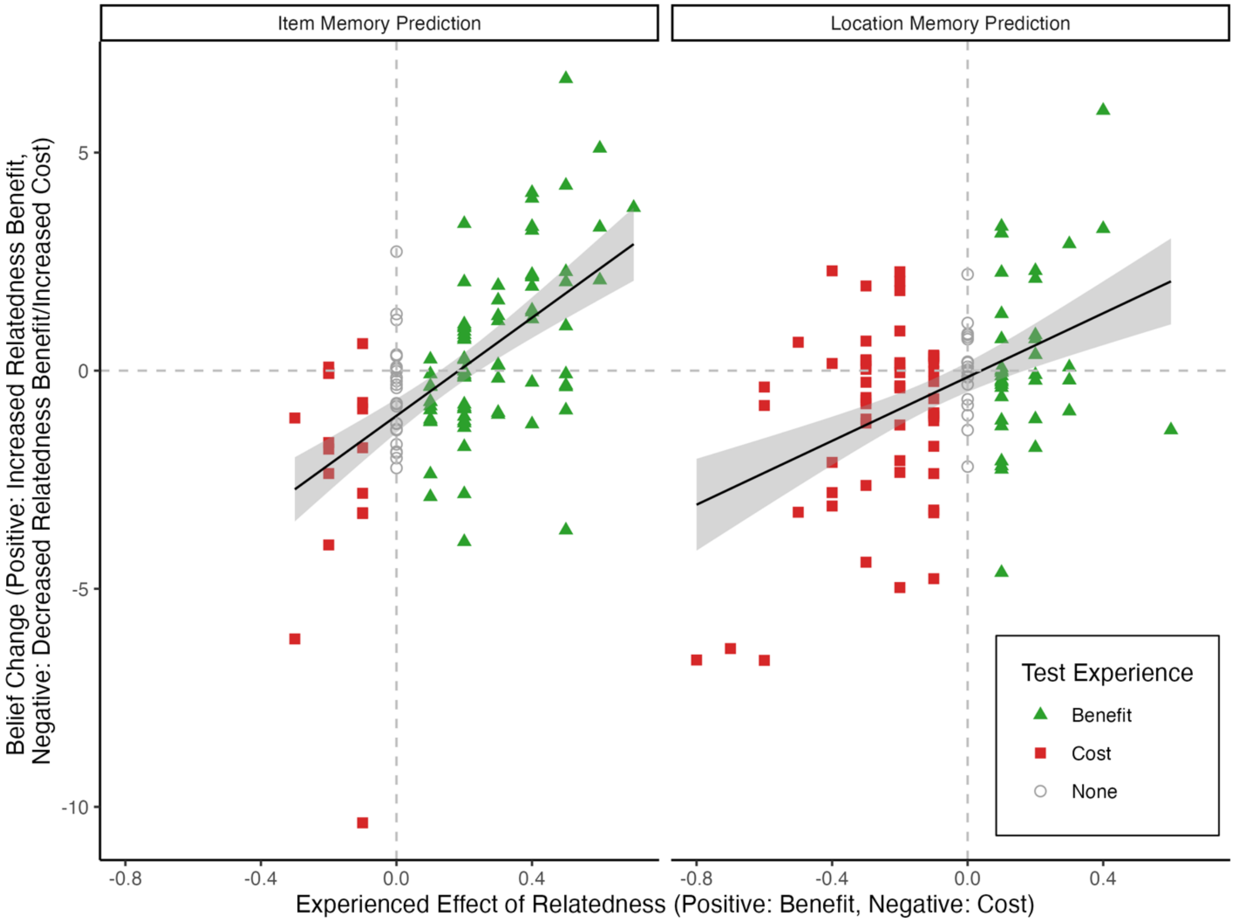


Figure 11. Experiment 9: Plot showing the relation between participants' objective related-unrelated performance difference, and their subjective related-unrelated belief change from pre-task to post-task. Shapes in the background represent individual participants' belief change as a

function of test experience (green triangles represent participants who performed better for the related display; red squares represent participants who performed better for the unrelated display; hollow grey circles represent participants who performed equally in both displays).

What Contributes to Participants' Final Predictions?

For both item and location memory, there was a positive relation between objective relatedness experiences (related – unrelated memory performance) and subjective belief change in the benefit of relatedness (change in related – unrelated predictions from pre-task to post-task). Subsequently, I wondered to what extent participants' final elicited beliefs were dependent on their initial beliefs as well as their experiences in the task.

To explore this question, I used linear regression to predict participants' post-task relatedness beliefs (related – unrelated). The predictors were their pre-task relatedness beliefs (related – unrelated) and their objective memory performance in the task (related – unrelated). Regressions were first performed separately for the item and location memory groups. The interaction terms were not significant (item: p = .162, location: p = .598) and hence not included in either model. For participants making item memory predictions, their final predictions were significantly influenced both by their initial predictions, b = 0.18 [95% CI: 0.04, 0.33], t = 2.51, p = .013, and by their objective memory performance, b = 4.34 [95% CI: 3.30, 5.38], t = 8.29, p < .001. A similar result was found for participants making location memory predictions: Their final predictions were significantly influenced both by their initial predictions, b = 0.23 [95% CI: 0.07, 0.39], t = 2.80, p = .006, and by their objective memory performance, b = 3.23 [95% CI: 2.19, 4.27], t = 6.13, p < .001. When I conducted a linear regression on the combined item and location data, I found the same pattern: no significant interaction terms (all ps > .170), but a significant effect of initial predictions, b = 0.20 [95% CI: 0.09, 0.31], t = 3.64, p < .001, as well as their objective memory performance, *b* = 3.75 [95% CI: 3.02, 4.48], *t* = 10.10, *p* < .001. After accounting for these effects, participants' final predictions did not differ for location vs. item, *b* = -0.25 [95% CI: -0.64, 0.14], *t* = 1.25, *p* = .211.

Self-reported post-task prediction reasoning

Participants' self-reported reasoning for their post-task predictions were coded as belonging to the various categories outlined in Table 25. The coding scheme was adapted from one that was previously developed (Lu & Risko, under review; see Chapter 2), with the addition of the task experience category. There were two independent coders who were blind to participants' predictions and condition assignment, and each reason could be assigned to multiple categories under the coding scheme. As a measure of inter-rater reliability, I calculated Cohen's kappa (κ) for each category, a measure which corrects for chance agreement between raters (Cohen, 1960). Cohen's κ ranged from 0.48 to 0.96, suggesting moderate to substantial agreement between the two raters (Cohen, 1960). Table 26 shows the percentage of responses classified into each category.

Table 25. Coding scheme for participant reasoning in Experiments 9 and 10

Belief Classification	Description	Example
Mental Imagery, Visualization	Participant indicates that they can more easily form mental imagery or visualize the items for one of the displays	*When remembering items from the same category, it is easier to visualize them when trying to remember them*
Associations, Connections, Cueing	Participant indicates that they can more easily form links between items for one of the displays	*It would be easier to recall items associated in my mind already*
Distinctiveness, Interference	Participant indicates that they can more easily distinguish between the items for one of the displays	*I get the same category items jumbled up whereas different category items are easier to recall*
Unspecific Related Preference	Participant reiterates that they find it easier to remember related	*It's easier to remember words that are related to one another*

	words, but no particular reason provided	
Unspecific Unrelated Preference	Participant reiterates that they find it easier to remember unrelated words, but no particular reason provided	*I think I can better recall things from multiple categories rather than from one category*
Indifference	Participant indicates that the conditions will have little influence on memory	*I feel like it does matter if the category is the same or different*
Memory Ability	Participant indicates that their performance is dependent on their memory ability	*My memory is fairly good for images*
Past Experience	Participant makes reference to pre-experiment experiences to support their predictions	*I've never been good at memory or locating games like matching pairs*
Task Experience	Participant makes reference to their experiences during the experiment to support their predictions	*Based on how I did for the two tasks, I noticed I was way better at remembering items from the same category*
Belief Change	Participant indicates that their belief about the effects of relatedness changed from pre- to post-task	*I initially thought that I would find it easier to remember the ones from the same category, but after trying the task I think I actually did better when the items were from different categories*
Other Reason	Participant provides an unlisted reason for preferring either of the displays	*If they're from the same category I can't think of a way of remembering them, but if they're from different categories I make up a story involving all of the items*

Table 26. Experiment 9: Percentages of participants' self-reported reasoning behind their predictions. Note that responses could be classified into multiple reasons, and some were unclassifiable/unusable, hence columns do not sum to 100%.

	Item Memory (*N*=122)			**Location Memory (*N*=118)**		
	Related Benefit (*N*=246)	Related Cost (*N*=31)	No Difference (*N*=89)	Related Benefit (*N*=112)	Related Cost (*N*=85)	No Difference (*N*=157)
Imagery	9.3%	9.7%	7.9%	6.3%	2.4%	-
Association	38.6%	6.5%	15.7%	21.4%	3.5%	1.9%

Distinctiveness	2.0%	19.4%	1.1%	6.3%	27.1%	7.6%
Unexplained Related Preference	13.0%	3.2%	-	18.8%	2.4%	2.5%
Unexplained Unrelated Preference	0.8%	9.7%	1.1%	-	1.2%	1.3%
Indifference	0.8%	-	4.5%	8.0%	5.9%	23.6%
Memory Ability	8.9%	6.5%	6.7%	10.7%	5.9%	19.7%
Past Experience	1.6%	3.2%	4.5%	0.9%	-	3.2%
Task Experience	37.0%	45.2%	67.4%	49.1%	69.4%	59.2%
Belief Change	2.8%	9.7%	9.0%	3.6%	8.2%	4.5%
Other Reason	4.5%	9.7%	1.1%	1.8%	1.2%	1.9%

Discussion

At pre-task, participants predicted that related items would be better remembered than unrelated items, both for the item memory test and for the location memory test. This effect was moderated by task, such that the aggregate relatedness benefit was predicted to be smaller overall in the location memory task: More individuals predicted a relatedness benefit for the item memory task, whereas more individuals predicted a cost or no difference of relatedness for the location memory task. In terms of objective memory performance, participants remembered more related items compared to unrelated items in the item memory test (a relatedness benefit) but remembered fewer related item locations compared to unrelated locations in the location memory test (a relatedness cost). Experiment 9 thus replicated both the objective cost of relatedness in a location memory task (Lu et al., 2023), as well as the tendency for participants to believe in a relatedness benefit even in the location memory context where relatedness was harmful to performance (i.e., a 'relatedness halo'; Lu & Risko, under review).

The primary interest was in how participants' predictions might change after their first-hand experiences in the task. For item memory predictions, the relatedness benefit was significant and equivalently sized across both the pre-task and post-task prediction timepoints. For location memory predictions, the relatedness benefit was significant only at the pre-task timepoint; it was no longer significant at the post-task timepoint. Since relatedness was associated with an objective benefit in the item memory test but an objective cost in the location memory test, this suggests that the changes I observed might be driven by performance differences in the two tests. In support of this notion, I found that an individual's objective relatedness performance in the task (i.e., related vs. unrelated performance difference) was a significant predictor of changes in relatedness predictions: Participants who experienced an objective cost of relatedness tended to move their predictions in a negative direction (i.e., shift away from a benefit and toward a cost) whereas those who experienced an objective benefit of relatedness were more likely to move in a positive direction (i.e., shift toward a greater benefit). This suggests that individuals were relying on their performance in the two conditions as a source of information for their predictions. The strength of this relation was not moderated by task; that is, an individual experiencing an objective relatedness benefit or cost (in either task) tended to shift their relatedness predictions accordingly. Although the test type influenced the likelihood of experiencing a relatedness benefit (more likely for item memory) or a cost (more likely for location memory), the influence of the experience once obtained was similar (Figure 11).

Finally, participants' post-task predictions about relatedness were a function both of their pre-task relatedness predictions and of their objective relatedness performance, and these effects were not moderated by task. Participants' objective relatedness performance (and their estimates of this performance) was correlated with their final predictions, which further suggested its importance as an informational base. Even after accounting for objective relatedness performance, however, pre-task relatedness predictions still positively predicted post-task relatedness predictions for both tasks. In other words, even as participants are partly basing their final judgments on their objective performance, they do not seem to completely abandon their initial belief. Final expressed beliefs about relatedness were predicted by both their initial predictions about relatedness and their recent experience in the task. This could have resulted from participants drawing upon their previous judgments directly, or from the indirect influence of the retrieved ideas and information that those previous judgments were based upon, or from a combination of the two routes of influence.

The results of Experiment 9 can be understood within the beliefs as constructive judgments framework (Lu & Risko, under review), wherein metamemory belief reports are understood to be dynamically constructed in response to specific elicitation contexts. When participants are queried for these beliefs, their responses are constructed from information that can be retrieved from memory at the time of the judgment (Cavanaugh et al., 1998). The experimental vignette and prompt consist of various cues that guide the retrieval of different kinds of information and lead to the activation of different naïve theories of memory (Lu & Risko, under review). At the pre-task timepoint, upon reading the task descriptions, people are more likely to activate ideas that are consistent with relatedness leading to a "benefit" (e.g., relatedness helps one make associations between items) instead of a "cost" (e.g., relatedness reduces item distinctiveness), leading to the initial aggregate prediction of a benefit. At the post-task timepoint, people have another source of information: how well they (think that they) had performed in the task.

Within the framework, test performance is likely to be readily retrieved from memory and drawn upon as it is both recent and presumably highly relevant to the prediction context (i.e., they are told to predict their performance in a future task that will be like the one they had experienced). In support of this notion, participants' self-reported reasoning (Table 5) suggests that their first-hand experience was the most dominant and informative cue for their post-task predictions. Nevertheless, people continued to invoke notions like associations (to explain benefit) and distinctiveness (to explain cost), which they may have initially retrieved in response to the judgment context (see Lu & Risko, under review) or formed during their experiences in the task. The framework predicts that participants' final relatedness beliefs, as expressed at post-task, are constructed from various sources of retrieved information, including their previous predictions – or the initial information that shaped these predictions, such as naïve theories that were cued by the context – and their retrieved performance in the task itself.

Finally, I note one unexpected finding: Item memory pre-task relatedness predictions were negatively correlated with their objective relatedness performance and estimates of that performance. I speculate that participants who believed that relatedness was beneficial may have subsequently spent less study effort on the related condition (in a self-defeating prophecy), which then had the ironic effect of decreasing the benefit. I did not observe this negative correlation in the location memory group, but there were also fewer people anticipating a relatedness benefit for this task. I previously reported that participants do spend less time studying the related items, even for the location memory test (Lu et al., 2023), but did not find this to be the main driver of the relatedness cost in location memory. Regardless of the cause, this could explain the non-significant bivariate correlation between pre-task and post-task relatedness predictions for item memory (Table 23).

Experiment 10

In Experiment 9, I observed a decrease from pre-task to post-task in participants' predictions of the relatedness benefit in the location memory task. This change was observed after the participants had studied both the related and unrelated item displays (study experience) and had performed their respective location memory tasks (test experience). Thus, it is unclear what aspects of participants' experiences contributed to this change, as they experienced both the study phase of the related and unrelated displays as well as the test phase of the related and unrelated displays before making post-task predictions. While I found that participants' post-task belief reports about relatedness were influenced by both their pre-task predictions and their retrieved performance in the task itself, this does not rule out the possible influence of study experiences on their final predictions. The aim of Experiment 10 was to disambiguate to what extent the change in participants' relatedness belief reports was driven by their study and/or test experiences.

One way to examine the nature of this change is by soliciting participants' predictions at different stages of the experiment, such as before and after presenting the study material and/or before and after the memory test (see also Koriat & Bjork, 2006a). Thus, in Experiment 10 I solicited participants' predictions at three stages of the experiment: pre-task, post-study, and post-test. To facilitate this, I employed a study-study-test-test design instead of the two separate study-test blocks design of Experiment 1 (study-test-study-test), such that participants in Experiment 2 studied both displays before being tested on both displays. Post-study predictions were elicited after the study phase (i.e., after participants had studied both the related and unrelated displays), and post-test predictions were elicited after the test phase (i.e., after participants had been tested on both displays).

By comparing participants' post-study predictions with their pre-task predictions, we can examine whether the study phase led to any changes in their relatedness belief reports. By comparing participants' post-test predictions with their pre-task predictions, we can investigate whether the test phase is necessary to drive changes in these beliefs. Previous studies have shown that test experience is a key factor in ‘mending metacognitive illusions’ (e.g., Benjamin, 2003; Koriat & Bjork, 2006a, 2006b; Yan et al., 2016). When there is a mismatch between participants’ initial predictions of memory and their actual memory, this can be corrected to some extent after experiencing the memory test. Exposure to the test reveals the disconnect between what participants initially thought would help memory and the reality of the test conditions. The results of Experiment 9 also supported this idea: Participants’ objective relatedness performance in the test was predictive of their final predictions about relatedness, and the degree of their belief change also depended on their objective test performance. Thus, I was rather confident in anticipating that test experience would influence participants’ final predictions. The contribution of study experience was, however, less clear. To address this question, I used participants’ pre-task predictions, post-study predictions, and objective performance to model their final predictions.

Method

Participants. Data from 165 participants were analyzed (120 women, 38 men, 4 other, 3 unknown, M = 20.11 years, SD = 3.65). The sample size was determined using a power analysis using G*Power (Faul et al., 2007) with a predicted pre-task relatedness difference of d = .22 (Lu and Risko, under review) to achieve a power of .80 at α = .05. Participants were University of Waterloo students who participated for course credit.

Procedure. The experimental procedure was similar to Experiment 9, with the following critical modification: Participants studied both displays—related and unrelated—consecutively before being tested on both displays consecutively. The order of the two study conditions was randomized for each participant, and test order was the same as study order: For example, a participant who studied the related display followed by the unrelated display would be tested in that order. This change also meant that participants experienced a single math distractor task between the study-study and test-test blocks. Finally, the postdiction tasks and self-reported reasoning tasks were omitted from Experiment 10. Figure 12 shows a schematic diagram of the experiment procedure.

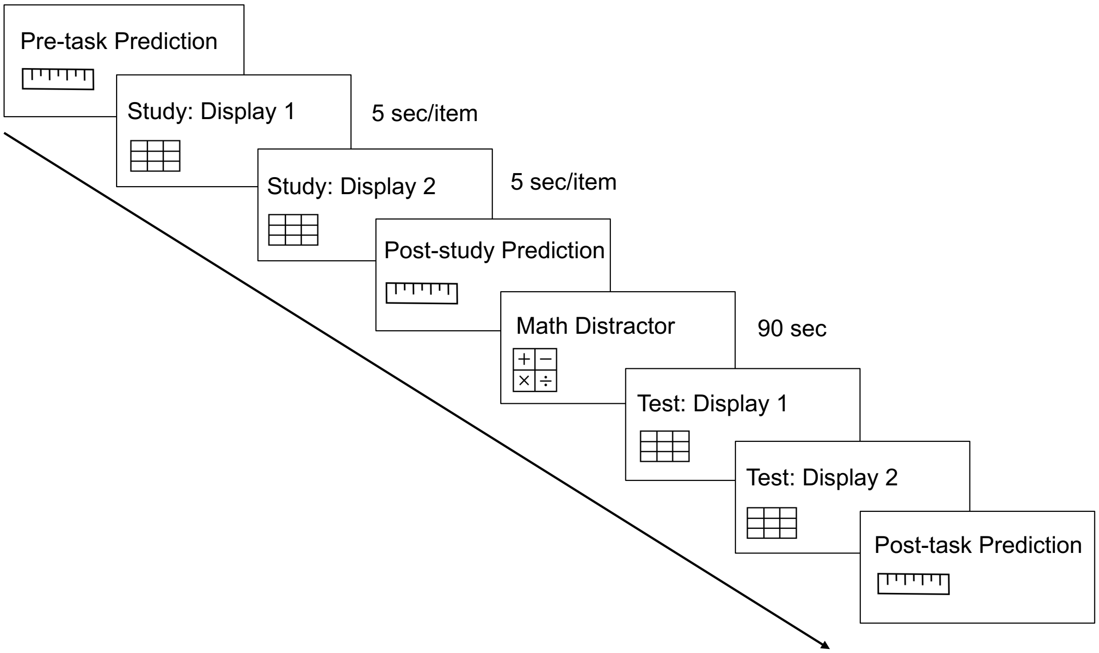


Figure 12. Schematic diagram of Experiment 10 procedure.

Results

Data from 64 participants were excluded according to the pre-registered exclusion criteria. They either (1) self-reported that they were not paying attention or did not give effort during the task; and/or (2) did not answer at least 14 math questions and achieve at least 70% correct in each math distractor task. The final sample size was *N* = 165 after replacement (related display first: *N* = 82, unrelated display first: *N* = 83).

Objective Memory Performance

A 2 (Relatedness: Related vs. Unrelated) x 2 (Block Order: Related-First vs. Unrelated-First) mixed-subjects ANOVA revealed a main effect of relatedness, *F*(1, 163) = 4.35, *MSE* = 0.03, *p* = .039, η_G^2 = .008, replicating the relatedness cost in objective memory (Related *M* = 3.16, *SD* = 2.25; Unrelated *M* = 3.55, *SD* = 2.13). There was no main effect of display order, *F*(1, 163) = 4.35, *MSE* = 0.07, *p* = .636, η_G^2 < .001, though there was a significant interaction of order with relatedness, *F*(1, 163) = 19.78, *MSE* = 0.03, *p* < .001, η_G^2 = .034. Follow-up analyses revealed that the relatedness cost was significant when the unrelated display came first, *F*(1, 82) = 21.36, *MSE* = 0.03, *p* < .001, η_G^2 = .065, but not when the related display came first, *F*(1, 81) = 2.79, *MSE* = 0.03, *p* = .099, η_G^2 = .011. Figure 13 depicts participants' objective location memory performance as a function of condition and display order.

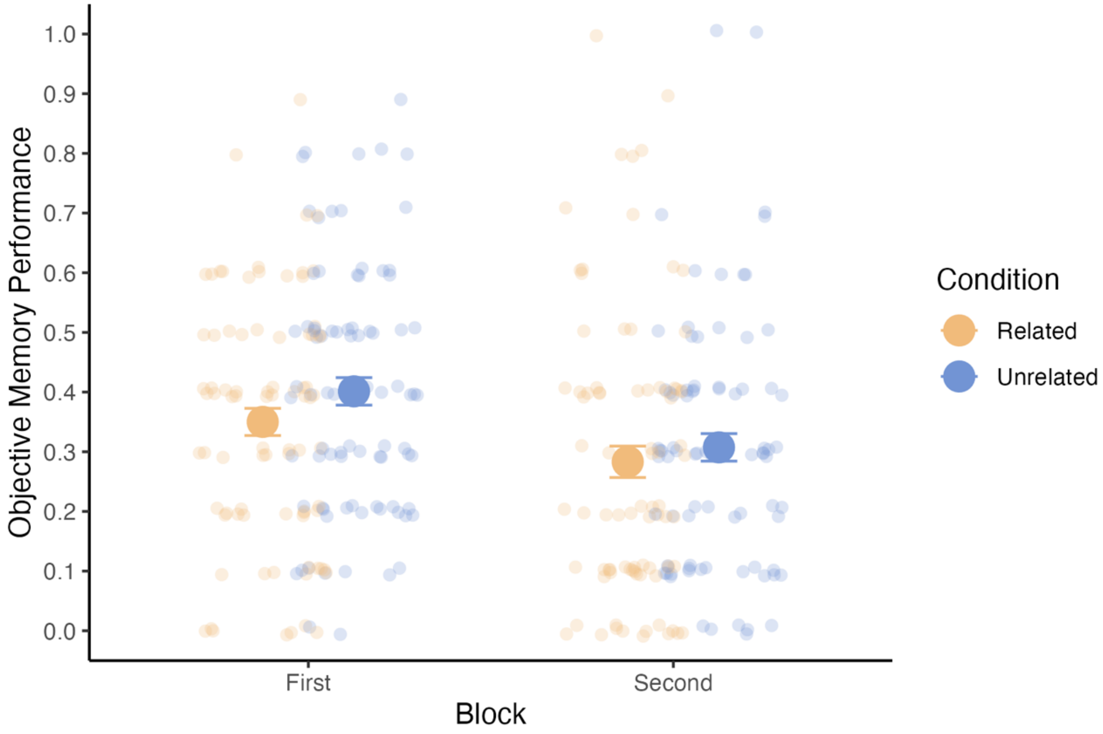


Figure 13. Experiment 10: Mean correct locations selected in the related and unrelated conditions, separated by block. Circles in the background represent individual participants' memory performance; error bars are standard errors.

Subjective Memory Predictions

Location Memory Predictions (Omnibus Analysis). A 2 (Relatedness: Related vs. Unrelated) x 3 (Prediction Time: Pre-task vs. Post-study vs Post-test) repeated measures ANOVA revealed a main effect of relatedness, $F(1, 164) = 17.81$, $MSE = 3.03$, $p < .001$, $\eta_G^2 = .014$, indicating that participants provided higher predictions for the related display compared to the unrelated display. There was also a main effect of prediction time, $F(1.80, 294.82) = 129.55$, $MSE = 3.35$, $p < .001$, $\eta_G^2 = .170$, indicating that participants' predictions decreased over the course of the experiment. Importantly, there was a significant interaction between relatedness and prediction time, $F(1.89, 310.20) = 5.15$, $MSE = 1.32$, $p = .007$, $\eta^2_G < .01$, indicating that the predicted difference between the related and unrelated displays depended on the prediction time-point.

As pre-registered, I followed up by performing two 2 x 2 ANOVAs (Relatedness x Pre-task vs. Post-study, and Relatedness x Pre-task vs. Post-test) as well as t-tests (related vs. unrelated) at each of the three time points.

Location Memory Predictions (Pre-task vs. Post-study). A 2 (Relatedness: Related vs Unrelated) x 2 (Prediction Time: Pre-task vs. Post-study) repeated measures ANOVA revealed a main effect of relatedness, $F(1, 164) = 22.57$, $MSE = 2.74$, $p < .001$, $\eta_G^2 = .024$, indicating that participants provided higher memory predictions for the related display. There was also a main effect of prediction time, $F(1, 164) = 72.40$, $MSE = 3.25$, $p < .001$, $\eta_G^2 = .087$, indicating that participants provided lower predictions at post-study compared to pre-task. Importantly, there was no significant interaction between relatedness and prediction time, $F(1, 164) = 1.63$, $MSE = 1.49$, $p = .203$, $\eta_G^2 < .001$, suggesting that the predicted relatedness difference did not change from pre-task to post-study.

Location Memory Predictions (Pre-task vs. Post-test). A 2 (Relatedness: Related vs Unrelated) x 2 (Prediction Time: Pre-task vs. Post-test) repeated measures ANOVA again revealed a main effect of relatedness, $F(1, 164) = 14.47$, $MSE = 2.36$, $p < .001$, $\eta_G^2 = .013$, as well as a main effect of prediction time, $F(1, 164) = 207.89$, $MSE = 3.74$, $p < .001$, $\eta_G^2 = .232$. Critically, there was a significant interaction, $F(1, 164) = 9.89$, $MSE = 1.30$, $p = .002$, $\eta_G^2 = .005$, suggesting that the predicted relatedness difference changed from pre-task to post-test.

I further conducted an exploratory 2 (Relatedness: Related vs Unrelated) x 2 (Prediction Time: Post-study vs. Post-test) repeated measures ANOVA and found a significant interaction, $F(1, 164) = 4.22$, $MSE = 0.97$, $p = .042$, $\eta_G^2 = .002$, that is, the predicted relatedness difference decreased from post-study to post-test. This further suggests that it was the test experience—and not the study experience— that drove the decrease in the predicted relatedness benefit observed at post-test.

I followed up these analyses with pre-registered paired sample t-tests (related vs. unrelated predictions) for each prediction time point (pre-task, post-study, and post-test). At the pre-task time point, participants predicted a relatedness benefit, $t(164) = 4.34$, $p < .001$, $d_z = 0.34$ [95% CI: 0.18, 0.49]. They continued to predict a relatedness benefit at the post-study time point, $t(164) = 3.25$, $p = .001$, $d_z = 0.25$ [95% CI: 0.10, 0.41]. However, there was no significant difference in their related and unrelated predictions at the post-test time point, $t(164) = 1.39$, $p = .164$, $d_z = 0.11$ [95% CI: -0.04, 0.26]. Table 27 shows participants' predictions at each time point.

Table 27. Experiment 10: Participants' mean (SD) metamemory predictions

	Related	**Unrelated**
Pre-task Prediction	5.43 (1.88)	4.70 (1.97)
Post-study Prediction	4.12 (2.03)	3.63 (1.86)
Post-test Prediction	2.99 (2.13)	2.81 (1.91)

Changes in relatedness belief direction. As in Experiment 9, participants were sorted into three groups based on their related – unrelated prediction difference (related benefit, related cost, and no difference). Table 28 shows the proportion of participants in each group.

An exploratory analysis was performed to quantify whether the proportion of participants expressing each kind of relatedness belief changed over time. I used generalized estimating equations to model beliefs about relatedness as an ordinal multinomial response (benefit > no difference > cost). Prediction time (pre-task, post-study, post-test) was entered in the model as the predictor with pre-task as the reference level, and subject was the clustering variable. I found that the proportion of participants in each group was not significantly different at post-study compared to pre-task, $b = 0.17$, 95% CI [-0.16, 0.50], $z = 1.00$, $p = .317$. However, the proportion of participants did change significantly at post-test compared to pre-task, $b = 0.64$, 95% CI [0.26, 1.02], $z = 3.33$, $p = .001$. When the reference level was changed to the post-study timepoint, I found that the proportion of participants in each group changed significantly from post-study to post-test, $b = 0.48$, 95% CI [0.18, 0.78], $z = 3.11$, $p < .002$, but there was no change from post-study to pre-task, $b = -0.15$, 95% CI [-0.48, 0.18], $z = 0.88$, $p = .380$, again suggesting that the observed changes were primarily driven by test experience rather than study experience.

Table 28. Experiment 10: Proportions of participants who predicted a related benefit, related cost, or no difference at the pre-task, post-study, and post-test timepoints

	Related Benefit	**Related Cost**	**No Difference**
Pre-task Prediction	57.6%	23.0%	19.4%
Post-study Prediction	52.1%	23.6%	24.2%
Post-test Prediction	37.0%	29.1%	33.9%

Relation between Objective Memory and Subjective Memory

Correlations between objective and subjective memory. Table 29 shows the bivariate correlations between participants' related – unrelated performance difference, and their related – unrelated predictions at pre-task, post-study, and post-test.

Table 29. Experiment 10: Bivariate correlations

	Objective Performance	**Pre-task Prediction**	**Post-study Prediction**	**Post-test Prediction**
Objective Performance	-			
Pre-task Prediction	.05	-		
Post-study Prediction	.22 **	.30 ***	-	
Post-test Prediction	.32 ***	.30 ***	.40 ***	-

Note. * p < .05, ** p < .01, *** p < .001

Objective Memory and Subjective Belief Change. Similar to Experiment 9, I performed linear regression using participants' related vs. unrelated performance differences in the memory task to predict changes in their subjective memory ratings for the related vs. unrelated conditions.[10] Again, I found a positive relation between participants' experienced objective memory performance difference (related – unrelated memory performance) and their subjective

[10] One extreme value was removed after visual inspection of the regression plot revealed this outlier.

belief change in the benefit of relatedness (change in related – unrelated predictions from pre-task to post-test), *b* = 1.69 [95% CI: 0.37, 3.01], *p* = .012. In other words, a participant who experienced a relatedness cost in the memory task tended to shift their belief in that direction whereas a participant who experienced a relatedness benefit tended to shift their belief toward a greater benefit. Figure 14 illustrates the relation between objective memory performance and subjective belief report change.

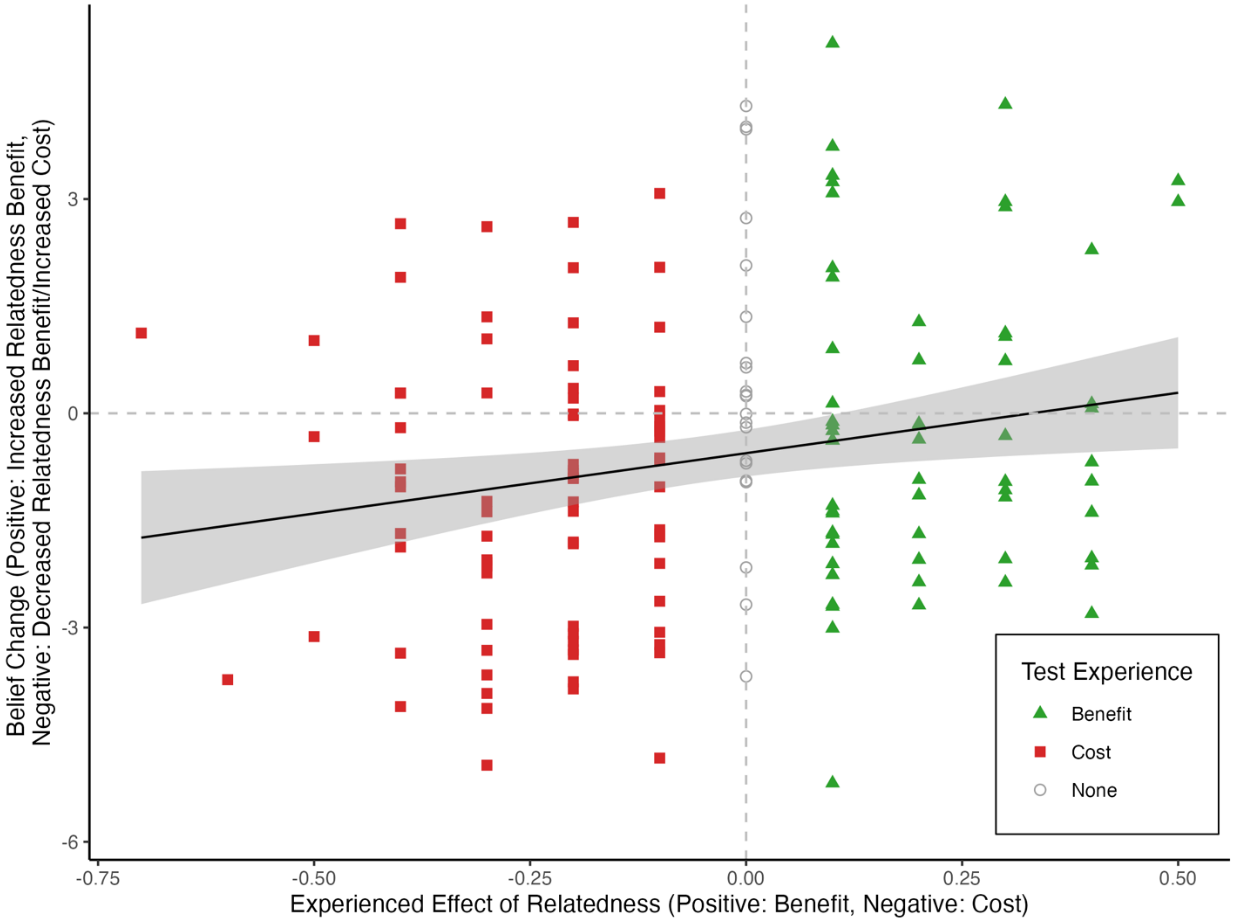


Figure 14. Experiment 10: Plot showing the relation between participants' objective related-unrelated performance difference, and their subjective related-unrelated belief change from pre-task to post-task. Shapes in background represent individual participants' belief change as a function of test experience (green triangles represent participants who performed better for the

related display; red squares represent participants who performed better for the unrelated display; hollow grey circles represent participants who performed equally in both displays).

What Contributes to Participants' Final Predictions?

I explored to what extent participants' final elicited beliefs were dependent on their previously reported beliefs as well as their experiences in the task. This was accomplished using linear regression to predict participants' post-test relatedness beliefs (related – unrelated). The predictors were their pre-task and post-study relatedness beliefs, as well as their objective memory performance in the task (related – unrelated). None of the interaction terms were significant (*p*s > .097) and hence were not included in the model. I found that all three predictors were significant: Pre-task relatedness predictions (*b* = 0.15 [95% CI: 0.05, 0.26], *t* = 2.88, *p* = .005), post-study relatedness predictions (*b* = 0.23 [95% CI: 0.11, 0.35], *t* = 3.82, *p* < .001), and objective relatedness performance (*b* = 1.62 [95% CI: 0.71, 2.53], *t* = 3.53, *p* = .001) all predicted participants' final relatedness predictions.

Discussion

In Experiment 10, I replicated both the objective relatedness cost in participants' location memory performance as well as the predicted relatedness benefit in their pre-task location memory predictions (Lu et al., 2023; Lu & Risko, under review). Similar to Experiment 9, I again observed that the predicted relatedness benefit decreased after participants had first-hand experience with the task. Critically, this decrease was only significant after test experience and not after study experience. The predicted relatedness benefit remained significant at the pre-task and post-study timepoints but was non-significant at the post-test timepoint. The proportions of participants expressing each kind of belief (benefit, cost, or no difference) was similar at pre-task and post-study, with a significant change occurring only at the post-test timepoint. Taken together, these results strongly suggest that test experience—not study experience—was the main driver of change in participants' relatedness beliefs.

Like Experiment 9, I again observed that individual participants shifted their predictions in line with their own test performance: Participants who experienced an objective cost of relatedness tended to shift toward predicting a cost (or a reduced benefit) whereas those who experienced an objective benefit of relatedness tended to shift toward predicting a greater benefit. There was also a positive correlation between objective relatedness performance and their post-test relatedness predictions, again suggesting that participants were using their test performance as an important piece of information for their final predictions. These results agree with previous studies showing that test experience is a key factor in 'mending metacognitive illusions' (e.g., Benjamin, 2003; Koriat & Bjork, 2006a, 2006b; Yan et al., 2016): The mismatch between participants' initial predictions of memory and their actual memory was improved after experiencing the memory test, presumably because the test reveals the disconnect between what participants initially thought would help memory and the reality of the test conditions.

I did find, however, that final relatedness predictions were a function of pre-task predictions and post-study predictions as well as objective performance. That is, even after accounting for participants' objective relatedness performance, both pre-task and post-study relatedness predictions continued to exert an influence on post-task relatedness predictions. While their final judgments were in part based on their actual test performance, their initial predictions remained influential, which aligns with the results of Experiment 1. After accounting for initial predictions, post-study predictions also contributed to final predictions.

Since the post-study prediction occurred after the study phase but before the test phase, it can be taken as a measure of how participants believed they would perform on the upcoming test. It is possible that post-study judgments partly reflect subjective experiences during study; however, it is important to note that I did not have any direct measure of in-the-moment study experiences (e.g., processing speed/fluency, subjective feelings of ease). An alternative interpretation (that does not preclude the first) is that since the post-study prediction only occurred after both study blocks had been completed, participants were somewhat removed from their initial encoding experiences, and the post-study prediction is reflective of their attempts to simulate retrieval at test (akin to a delayed judgment of learning effect; Dunlosky & Nelson, 1991; Rhodes & Tauber, 2011). Either interpretation notwithstanding, these results suggest that participants drew upon their post-study judgments during the post-task judgment phase and/or accessed the information that the post-study judgments themselves were based upon.

Within the beliefs as constructive judgments framework (Lu & Risko, under review), I propose that participants drew on all these sources of information to some extent – their pre-task judgment (and/or associated information), post-study judgment (and/or associated information), and perceived test performance – when queried for their final beliefs. The framework posits that each judgment is in turn based on the available and accessible information at any given time. Initially, at the pre-task timepoint, participants responded on the basis of the information that the vignette and prompt cued them to retrieve. At the post-study timepoint, participants could retrieve their initial judgments (and associated information), their experiences during the study phase, and/or the results of any retrieval attempts. These results suggested that while participants' post-study beliefs about relatedness did not decrease significantly from their pre-task beliefs, the correlation between the two belief reports was only moderate ($r = .30$), so participants were presumably not relying entirely on their pre-task judgments at that point.

Finally, at the post-test timepoint, participants' immediate test experiences were clearly influential, as these were found to be driving the observed decrease in relatedness beliefs. This finding is consistent with the notion that test performance is highly recent and diagnostic information, thus contributing substantially to final belief reports. However, even after accounting for the contribution of this information, participants' previous two judgments (pre-task and post-study) were still influential to some degree. As noted earlier, they might be retrieving their previous judgments directly, or the information that influenced those judgments, or both. At each timepoint, I assume that each participant draws on some combination of the informational sources that are accessible to them.

Experiment 11

In both of the previous experiments, I observed a decrease in participants' predictions of the relatedness benefit (in the location memory task) that appeared to be primarily driven by their test performance: The decrease was only significant after the test phase, and the extent to which participants changed their relatedness beliefs was dependent upon their individual performance in the test. In Experiment 9, I found that an individual's objective performance was associated with changes in belief for participants in both the item memory group and the location memory group; however, since the item memory group usually experienced a benefit whereas the location memory group usually experienced a cost, only the location group exhibited a decrease in the relatedness benefit overall. These results suggested that the relation between objective performance and belief change may have been *stronger* for the item memory group. One possible reason for this observation was that participants taking the item recall test may have had more insight into their performance compared to those taking the location memory test: In the absence of feedback, an item memory participant would still be able to see that they recalled 8 related items compared to 6 unrelated items, whereas the results of a location memory test are more opaque. In support of this hypothesis, Table 3 suggests that participants in the item memory condition were better at estimating their true level of performance (postdiction-performance correlation of .66 vs. only .47 for location memory).

According to the beliefs as constructive judgments framework, the test information contributes significantly to participants' final (post-task) relatedness beliefs because it is highly diagnostic and accessible. Following this logic, a manipulation that increases a piece of information's diagnosticity and/or accessibility at the post-test timepoint should increase the relative contribution of that piece of information. Thus, in Experiment 3 half of the participants were given direct test feedback, which I anticipated would increase the diagnosticity and accessibility of the test information. I anticipated that the presence of feedback would increase participants' reliance on the test information for their post-task judgment and that this would lead them to be more likely to decrease their relatedness beliefs after test experience. I predicted that participants who received feedback would thus show a stronger relation between their test performance and their final predictions. I also wondered whether this increased reliance on the test information would lead to a corresponding decrease in the reliance on other information (e.g., pre-task predictions) for participants receiving feedback.

Another goal of Experiment 11 was to examine how participants' subjective confidence in their relatedness predictions (*belief credence*) would change after experience with the task. Lu and Risko (under review) reported that, prior to any task experience, people who predicted a benefit of relatedness were more confident than were people who predicted a cost or no difference. According to the self-consistency model of subjective confidence (Koriat, 2012; see also Koriat, 2013; 2018; 2024), confidence tracks the extent to which our chosen answer is consistent across retrieval attempts. If most of the retrieved information points toward the same answer, then this is expressed as high confidence; conversely, if the information is contradictory or supports different answers, then confidence will be lower. Lu and Risko (under review) speculated that, given the task vignette and prompt, information that supported a relatedness benefit may have been more consistently cued and retrieved than information supporting a cost, thus participants who expressed a benefit were also more confident. In Experiment 11, I examined how participants' credence in their relatedness beliefs might change after they had acquired task experience as an additional information source.

Method

Participants. Data from 400 participants were analyzed (195 women, 194 men, 4 other, 7 unknown, *M* = 41.16 years, *SD* = 12.58). The sample size was determined using a power analysis using G*Power (Faul et al., 2007) to achieve .80 power to compare correlations of .47 vs. .66 (from Experiment 9; the post-task prediction-performance correlation). Participants were recruited from Prolific and paid GB £1.25 for 10 minutes.

Procedure. The experimental procedure was similar to the location memory condition in Experiment 9. The main manipulation was that half of the participants were randomly assigned to receive post-test performance feedback (*N* = 205) while the other half did not receive any feedback (*N* = 195). Other procedural differences will be described below. Figure 15 shows a schematic diagram of the experiment procedure.

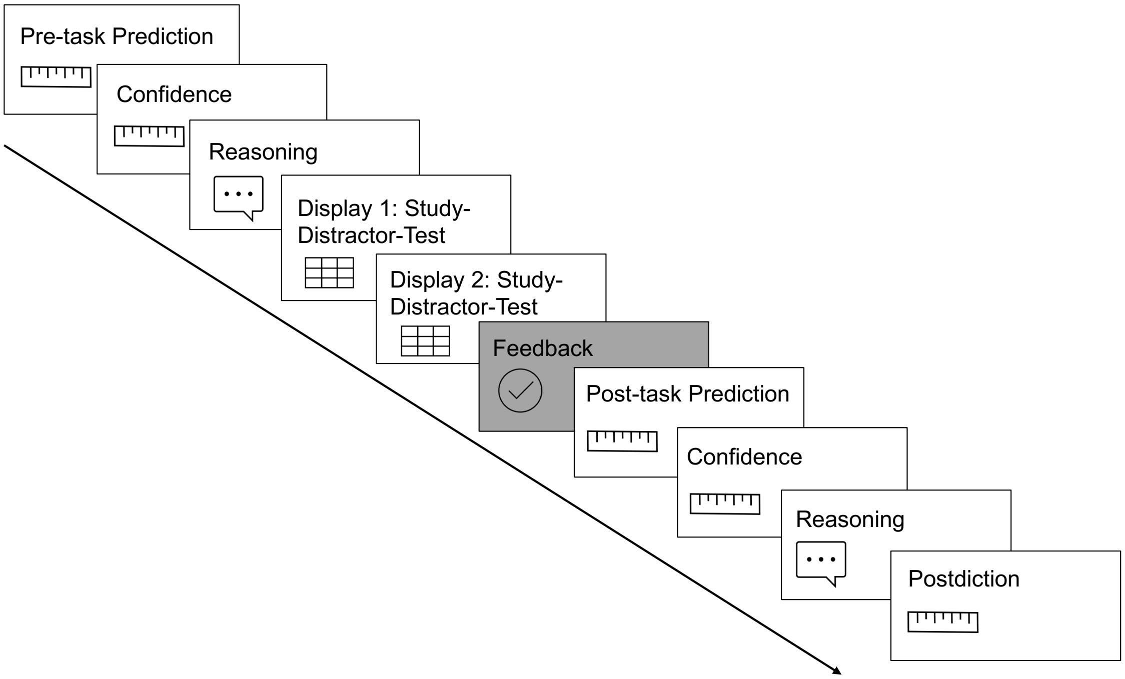


Figure 15. Schematic diagram of Experiment 11 procedure.

Pre-task Prediction. This was identical to Experiment 9.

Prediction Confidence. Participants were asked to provide a confidence rating for their pre-task predictions. They saw one of the following prompts depending on their pre-task predictions: "*You estimated that you would remember more items in the display where [all items are from the same category/where all items are from different categories], compared to the display where [all items are from different categories/all items are from the same category]"* if they predicted a benefit or a cost of relatedness, or "*You estimated that you would remember the same number of items in the display where all items are from the same category, and the display where all items are from different categories"* if they predicted the same number of locations for each type of display. They were asked to rate how confident they were that this was correct on a 1 (not at all confident) to 7 (extremely confident) scale.

Self-reported Reasoning. Participants were asked to explain the reasoning behind their predictions. The wording of the prompt was changed from Experiment 9, so that participants would be more likely to reason about relatedness (Lu & Risko, under review):

> "In the last task we asked you to make predictions about the number of words for which you would be able to correctly recall the location. (*Relatedness Benefit/Cost*: You estimated that you would remember more items in the display where all items are from the same category/where all items are from different categories, compared to the display where [where all items are from different categories/all items are from the same category]. *No Difference*: You estimated that you would remember same number of items in the display where all items are from the same category, and the display where all items are from different categories.) Please tell us about your reason(s) for this prediction. There is no right or wrong answer, as we're interested in how individuals make predictions about memory."

Encoding Task and Location Memory Tests. These were identical to Experiment 1.

Feedback Screen. If participants were in the feedback group, the following message was displayed:

> "You have completed this task. You scored [X] out of 10 correct locations for the first display ([items in the same category/items from different categories]). You scored [Y] out of 10 correct locations for the second display ([items in the same category/items from different categories])."

If participants were in the no feedback group, they were only told that they had completed the task. All participants then clicked on a 'Continue' button to proceed to the next screen.

Post-task Prediction. This was similar to the pre-task prediction #2 in Experiment 9 and post-test prediction in Experiment 10, in that participants were asked to make predictions about the two types of displays for a future experiment.

Comprehension Check #1. Participants were asked to respond to the following multiple-choice question: “On the previous page, you were to make estimates about: (i) your performance in an upcoming task, (ii) your performance in a previous task, or (ii) it did not specify.”

Prediction Confidence and Self-reported Reasoning #2. Participants were asked to provide a confidence rating for their post-task predictions, and to explain the reasons for these predictions. The prompts were identical to the pre-task.

Performance Postdiction. This was similar to the postdiction in Experiment 9, in that participants were asked to estimate their performance for the two types of displays in the memory task.

Comprehension Check #2. Participants were asked to respond to a comprehension check multiple choice question that was identical to the first.

Results

184 participants were excluded based on one or more of the pre-registered exclusion criteria. They either (1) self-reported that they were not paying attention or did not give effort during the task; and/or (2) did not answer at least 14 math questions and achieve at least 70% correct in either math distractor task; (3) failed to answer either of the final two comprehension check questions correctly. Of the 184 exclusions, 145 participants were only excluded due to failing the final comprehension checks (i.e., they possibly misunderstood or confused the post-task prediction and postdiction questions); I provide the pre-task prediction and location memory

analyses including these participants in Appendix C. All of the analyses below were pre-registered unless otherwise stated.

Objective Memory Predictions

One of the goals of Experiment 11 was to examine the replicability of the relatedness cost in location memory (Lu et al., 2023). Thus, I pre-registered mixed-effects analyses that used the same dependent measures as Lu et al. (2023): location accuracy, Euclidean distance, and item substitution rate, and found a significant relatedness cost in all. These detailed analyses can be found in Appendix C. To keep the focus of the current article on metamemory predictions, I only note here that I found a significant relatedness cost in objective location memory performance, $F(1, 399) = 28.96$, $MSE = 0.03$, $p < .001$, $\eta_G^2 = .022$, which was consistent with Experiments 9 and 10. Table 30 shows participants' memory performance and their postdictions for Experiment 11.

Manipulation Checks

I conducted two checks confirming the soundness of the feedback manipulation. First, I examined whether participants who received the feedback had more insight into their memory performance (i.e., their postdiction scores should be more accurate). Second, I confirmed that participants' location memory performance did not vary between the feedback conditions.

Feedback and Postdiction Accuracy. I conducted a 2 (Relatedness: Related vs. Unrelated) x 2 (Group: Feedback vs. No Feedback) ANOVA on participants' performance versus postdiction absolute deviation scores, which were calculated by taken the absolute difference (ignoring direction) between their true performance and their postdiction estimate. The effect of feedback was significant, $F(1, 398) = 26.68$, $MSE = 2.72$, $p < .001$, $\eta_G^2 = .044$, such that participants who received feedback reported postdiction scores that were closer to their actual performance. The effect of relatedness was not significant, $F(1, 398) = 0.64$, $MSE = 1.26$, $p = .425$, $\eta_G^2 < .001$, and neither was the interaction, $F(1, 398) = 0.18$, $MSE = 1.26$, $p = .668$, $\eta_G^2 < .001$.

Feedback and Location Memory Performance. I conducted a 2 (Relatedness: Related vs. Unrelated) x 2 (Group: Feedback vs. No Feedback) ANOVA on participants' location memory performance. The effect of feedback was not significant, $F(1, 398) = 0.63$, $MSE = 0.08$, $p = .429$, $\eta_G^2 = .001$, and neither was the interaction, $F(1, 398) = 0.04$, $MSE = 0.03$, $p = .848$, $\eta_G^2 < .001$. The effect of relatedness was significant, $F(1, 398) = 28.92$, $MSE = 0.03$, $p < .001$, $\eta_G^2 = .022$, such that the related items continued to be associated with worse performance than the unrelated items.

Subjective Memory Predictions

Table 31 shows participants' memory predictions at pre-task and post-task for Experiment 11.

Effect of Prediction Time, Relatedness, and Feedback. I conducted a 2 (Prediction Time: Pre-task vs. Post-task) x 2 (Relatedness: Related vs. Unrelated) x 2 (Group: Feedback vs. No Feedback) ANOVA on participants' location memory predictions. There was a main effect of relatedness, indicative of a predicted relatedness benefit, $F(1, 398) = 15.43$, $MSE = 1.53$, $p < .001$, $\eta_G^2 = .004$. There was also a main effect of prediction time, such that predictions decreased from pre-task to post-task, $F(1, 398) = 119.30$, $MSE = 3.58$, $p < .001$, $\eta_G^2 = .069$. The main effect of feedback was not significant, $F(1, 398) = 2.34$, $MSE = 8.17$, $p = .127$, $\eta_G^2 = .003$. The three-way interaction was not significant ($p = .761$). There was a significant interaction between prediction time and feedback, $F(1, 398) = 8.73$, $MSE = 3.58$, $p = .003$, $\eta_G^2 = .005$, as well as between prediction time and relatedness, $F(1, 398) = 35.72$, $MSE = 1.27$, $p < .001$, $\eta_G^2 = .008$.

The interaction between feedback and relatedness was not significant, $F(1, 398) = 0.94$, $MSE = 1.53$, $p = .334$, $\eta_G^2 < .001$. I follow up on both of the significant two-way interactions below.

Effect of Prediction Time Depending on Feedback. The significant interaction between prediction time and feedback suggested that feedback moderated participants' prediction changes across time. I conducted a one-way ANOVA on the effect of prediction time (Pre-task vs. Post-task) for each feedback group. Participants in the no feedback group gave significantly lower predictions after task experience, $F(1, 194) = 97.18$, $MSE = 1.73$, $p < .001$, $\eta_G^2 = .135$. A decrease was also found for the feedback group, but to a lesser extent, $F(1, 204) = 31.55$, $MSE = 1.85$, $p < .001$, $\eta_G^2 = .044$.

Changing Effect of Relatedness Across Time. The significant interaction between prediction time and relatedness suggested that the predicted relatedness difference decreased from pre-task to post-task.[11] To follow-up, I conducted a one-way ANOVA on the effect of relatedness for each feedback group. Participants gave higher pre-task predictions for the related items display than for the unrelated items display, $F(1, 399) = 51.70$, $MSE = 1.29$, $p < .001$, $\eta_G^2 = .027$. At post-task, their related and unrelated predictions were not significantly different, $F(1, 399) = 1.20$, $MSE = 1.50$, $p = .274$, $\eta_G^2 < .001$.

Table 30. Experiment 11: Mean (SD) metamemory predictions

	Feedback (*N*=205)		**No Feedback (*N*=195)**	
	Related	**Unrelated**	**Related**	**Unrelated**
Pre-task Prediction	5.09 (1.77)	4.59 (1.87)	5.23 (1.68)	4.57 (1.64)
Post-task Prediction	4.02 (2.04)	4.16 (2.19)	3.56 (1.92)	3.62 (1.92)

[11] This is equivalent to the analysis that was pre-registered to replicate the relatedness belief change observed in Experiment 9, i.e., 2 (Pre-Task vs. Post-Task) x 2 (Related vs. Unrelated) ANOVA and follow-up t-tests.

Table 31. Experiment 11: Mean (SD) memory performance and postdictions

	Feedback (*N*=205)		**No Feedback (*N*=195)**	
	Related	**Unrelated**	**Related**	**Unrelated**
Performance	4.06 (2.40)	4.74 (2.32)	3.88 (2.22)	4.61 (2.55)
Postdiction	4.00 (2.25)	4.31 (2.31)	3.25 (1.75)	3.28 (2.07)

Changes in related belief direction across time. As in earlier experiments, participants were categorized according to whether they had predicted a benefit, a cost, or no difference. Table 32 shows the proportions of participants expressing each kind of belief. To quantify whether the proportion of participants expressing each type of relatedness belief changed after experience, a generalized estimating equation was used to model beliefs about relatedness as an ordinal multinomial response (benefit > no difference > cost). The two predictors were prediction time (pre-task vs. post-task) and presence of feedback (feedback vs. no feedback), and subject was the clustering variable. The interaction was not significant (p = .256), so it was removed from the model. I found that the proportion of participants in each group changed significantly from pre- to post-task, b = 0.81, 95% CI [0.56, 1.05], z = 6.32, p < .001, but that feedback did not have a significant effect, b = 0.17, 95% CI [-0.11, 0.44], z = 1.19, p = .233.

Table 32. Experiment 11: Proportions of participants who predicted a related benefit, related cost, or no difference

	Feedback (*N*=205)			**No Feedback (*N*=195)**		
	Related Benefit	**Related Cost**	**No Difference**	**Related Benefit**	**Related Cost**	**No Difference**
Pre-task Prediction	40.5%	16.4%	43.9%	47.7%	10.7%	41.0%
Post-task Prediction	28.3%	29.7%	43.4%	23.1%	21.0%	54.9%

Relation between Objective Memory and Subjective Memory

Tables 33 and 34 shows the bivariate correlations between participants' objective performance, predictions, and postdictions, separated by feedback.

Correlation between performance and post-task beliefs. The bivariate correlation of related vs. unrelated performance with post-task relatedness belief was found to be significant, *r*(398) = .49, *p* < .001. While this relation was significant in both the feedback group, *r*(203) = .58, *p* < .001, and the no feedback group, *r*(193) = .39, *p* < .001, a two-sided Fisher's *z* test revealed that it was stronger in the feedback group, *z* = 2.55, *p* = .011.

Correlation between pre- and post-task beliefs. The bivariate correlation of pre-task relatedness belief with post-task relatedness belief fell short of significance, *r*(398) = 0.09, *p* = .062. This relation was not significant in either the feedback group, *r*(203) = 0.06, *p* = .358, or the no feedback group, *r*(193) = 0.13, *p* = .064. A two-sided Fisher's *z* test further suggested that the correlations did not differ by feedback, *z* = 0.68, *p* = .494.

Table 33. Experiment 11: Bivariate correlations for feedback group

	Feedback (*N*=205)			
	Objective Performance	Pre-task Prediction	Post-task Prediction	Postdiction
Objective Performance	-			
Pre-task Prediction	.11	-		
Post-task Prediction	.58 ***	.06	-	
Postdiction	.61 ***	.08	.67 ***	-

Note. * *p* < .05, ** *p* < .01, *** *p* < .001

Table 34. Experiment 11: Bivariate correlations for no feedback group

	No Feedback (*N*=195)			
	Objective Performance	Pre-task Prediction	Post-task Prediction	Postdiction
Objective Performance	-			
Pre-task Prediction	.07	-		
Post-task Prediction	.39 ***	.13	-	
Postdiction	.43 ***	.12	.75 ***	-

Note. * *p* < .05, ** *p* < .01, *** *p* < .001

Objective Memory and Subjective Belief Change. I pre-registered a linear regression using participants' related vs. unrelated performance differences in the memory task (i.e., objective relatedness task experience) to predict changes in their subjective memory ratings for the related vs. unrelated conditions (i.e., subjective relatedness belief change). The key question of interest was whether there would be an interaction between feedback and task experience (i.e., if the effect of task experience on belief change depended on whether participants received feedback or not). The interaction of feedback and task experience was not significant, *b* = 1.24 [95% CI: -0.36, 2.85], *t* = 1.53, *p* = .128. Therefore, I report the model without the interaction. Task experience was significantly predictive of belief change, *b* = 2.68, [95% CI: 1.88, 3.48], *t* = 6.57, *p* < .001, while presence of feedback was not, *b* = 0.06, [95% CI: -0.37, 0.48], *t* = 0.26, *p* = .797. Figure 16 illustrates the relation between objective memory performance and subjective belief report change.

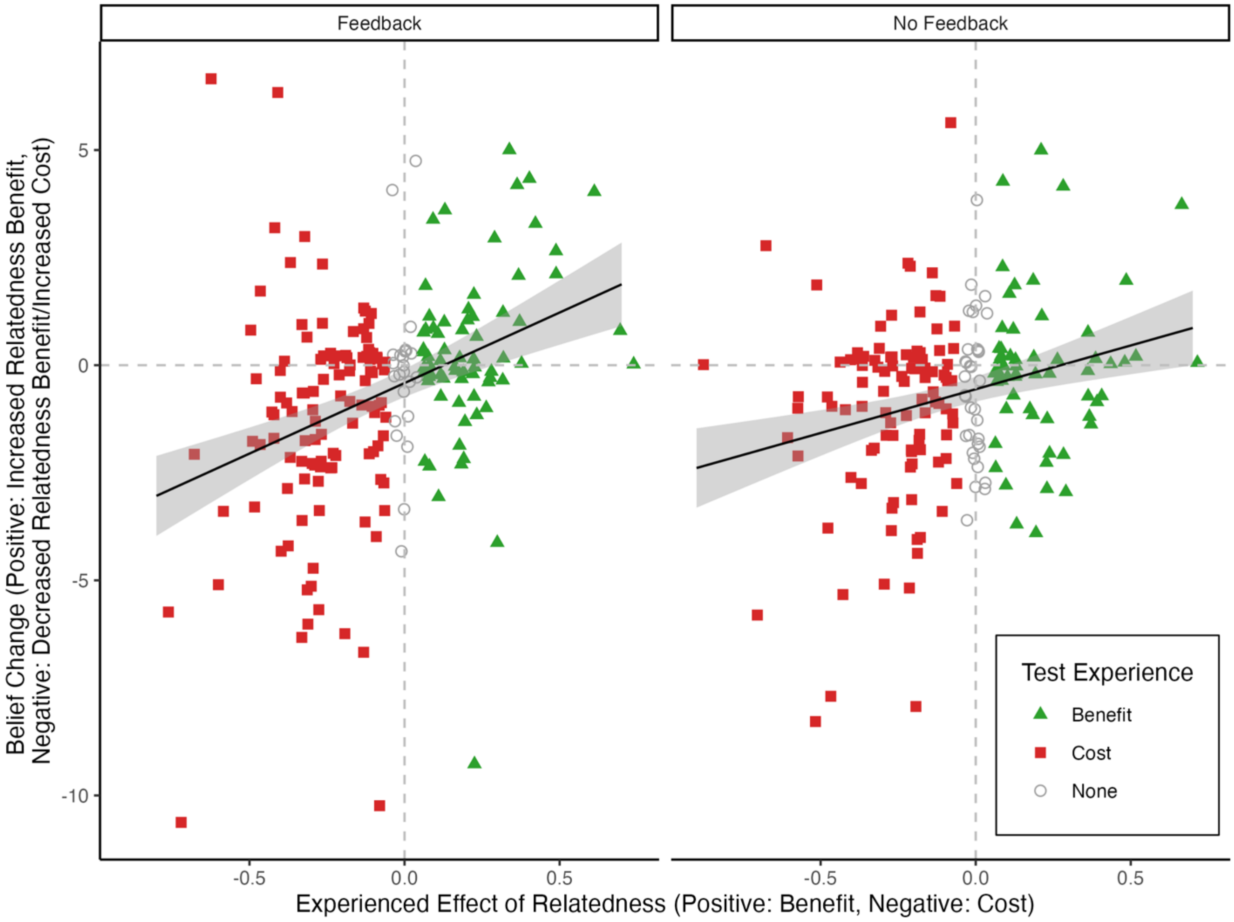


Figure 16. Experiment 11: Plot showing the relation between participants' objective related-unrelated performance difference, and their subjective related-unrelated belief change from pre-task to post-task. Shapes in background represent individual participants' belief change as a function of test experience (green triangles represent participants who performed better for the related display; red squares represent participants who performed better for the unrelated display; hollow grey circles represent participants who performed equally in both displays).

What Contributes to Participants' Final Predictions?

I pre-registered a linear regression analysis that examined to what extent participants' final elicited beliefs were dependent on their initial beliefs as well as on their experiences in the task (and whether feedback played a role). The outcome variable was participants' post-task relatedness beliefs (related – unrelated). The predictor variables were pre-task relatedness beliefs (related – unrelated), objective memory performance in the task (related – unrelated), and the presence of feedback; interaction terms were included if they significantly improved model fit. The final model included a significant interaction between objective memory performance and feedback, *b* = 1.59 [95% CI: 0.46, 2.71], *t* = 2.77, *p* = .006, suggesting that the effect of objective memory performance depended upon the presence of feedback. Estimates for the three main predictors were obtained from the model with the interaction removed: There was no effect of initial predictions, *b* = 0.05 [95% CI: -0.04, 0.15], *t* = 1.10, *p* = .272, and no effect of feedback, *b* = -0.09 [95% CI: -0.39, 0.20], *t* = 0.62, *p* = .538; only the effect of performance was significant, *b* = 3.21 [95% CI: 2.64, 3.78], *t* = 11.09, *p* < .001. I followed up with separate models for each feedback group. For participants who received feedback, their final predictions were significantly influenced by their objective memory performance, *b* = 4.01 [95% CI: 3.22, 4.79], *t* = 10.07, *p* < .001, but not by their initial predictions, *b* = 0.00 [95% CI: -0.12, 0.12], *t* = 0.01, *p* = .994. A similar result was found for participants who did not receive feedback: Their final predictions were significantly influenced by their objective memory performance (though to a lesser extent), *b* = 2.36 [95% CI: 1.54, 3.17], *t* = 5.71, *p* < .001, but not by their initial predictions, *b* = 0.12 [95% CI: -0.03, 0.27], *t* = 1.62, *p* = .107.

Pre- and post-task self-reported prediction reasoning

The same coding scheme was followed from Experiment 9: two independent coders were blind to participants' predictions and condition assignment, and each reason could be assigned to multiple categories under the coding scheme. Cohen's κ ranged from 0.76 to 0.90, suggesting substantial to almost perfect agreement between the raters, except for the other reason category which had moderate agreement (κ = 0.44). Tables 35 and 36 shows the percentage of responses classified into each category for pre-task and post-task predictions respectively.

Table 35. Experiment 11: Percentages of participants' self-reported reasoning behind their pre-task predictions. Note that responses could be classified into multiple reasons, and some were unclassifiable/unusable, hence columns do not sum to 100%.

Pre-task	**Feedback**			**No Feedback**		
	Related Benefit (*N*=83)	Related Cost (*N*=32)	No Difference (*N*=90)	Related Benefit (*N*=93)	Related Cost (*N*=22)	No Difference (*N*=80)
Imagery	4.8%	9.4%	-	8.6%	9.1%	-
Association	34.9%	3.1%	-	38.7%	-	-
Distinctiveness	-	71.9%	1.1%	1.1%	63.6%	-
Unexplained Related Preference	37.3%	-	1.1%	30.1%	-	3.8%
Unexplained Unrelated Preference	-	6.3%	1.1%	2.2%	13.6%	1.3%
Indifference	-	3.1%	44.4%	1.1%	-	47.5%
Memory Ability	13.3%	3.1%	38.9%	6.5%	13.6%	38.8%
Past Experience	1.2%	-	8.9%	-	3.8%	-
Other Reason	4.8%	6.3%	-	9.7%	4.5%	-

Table 36. Experiment 11: Percentages of participants' self-reported reasoning behind their post-task predictions. Note that responses could be classified into multiple reasons, and some were unclassifiable/unusable, hence columns do not sum to 100%.

Post-task	**Feedback**			**No Feedback**		
	Related Benefit (*N*=58)	Related Cost (*N*=58)	No Difference (*N*=89)	Related Benefit (*N*=45)	Related Cost (*N*=43)	No Difference (*N*=107)
Imagery	1.7%	3.4%	2.2%	4.4%	4.7%	0.9%
Association	15.5%	5.2%	5.6%	11.1%	2.3%	3.7%
Distinctiveness	6.9%	24.1%	4.5%	-	25.6%	1.9%
Unexplained Related Preference	8.6%	1.7%	1.1%	20.0%	2.3%	7.5%

Unexplained Unrelated Preference	1.7%	-	-	-	4.7%	-
Indifference	8.6%	10.3%	23.6%	8.9%	2.3%	31.8%
Memory Ability	3.4%	5.2%	20.2%	6.7%	-	16.8%
Past Experience	-	-	1.1%	-	-	-
Task Experience	58.6%	72.4%	56.2%	53.3%	79.1%	59.8%
Belief Change	19.0%	31.0%	11.2%	8.9%	41.9%	17.8%
Other Reason	3.4%	-	3.4%	8.9%	2.3%	0.9%

Prediction Confidence

Pre- and post-task prediction confidence. I pre-registered an analysis that examined how confident participants were in their relatedness predictions. A 2 (Prediction Time: Pre-task vs. Post-task) x 2 (Group: Feedback vs. No Feedback) ANOVA revealed that participants were less confident in their beliefs at the post-task time point, *F*(1, 398) = 63.63, *MSE* = 1.47, *p* < .001, η_G^2 = .047. The main effect of feedback was not significant, *F*(1, 398) = 2.57, *MSE* = 3.28, *p* = .110, η_G^2 = .004. These effects were qualified by a significant interaction, *F*(1, 398) = 4.50, *MSE* = 1.47, *p* = .034, η_G^2 = .003, so I followed up by examining the effect of prediction time for each feedback group. For participants who received no feedback, their prediction confidence significantly decreased from pre-task to post-task, *F*(1, 194) = 46.09, *MSE* = 1.59, *p* < .001, η_G^2 = .070. This decrease also occurred for participants who received feedback, though to a lesser extent, *F*(1, 204) = 19.01, *MSE* = 1.36, *p* < .001, η_G^2 = .028. Table 37 shows participants' prediction confidence (pre-task and post-task) as a function of feedback.

Table 37. Experiment 11: Participants' mean (SD) prediction confidence

	Feedback (*N*=205)	**No Feedback (*N*=195)**
Pre-task Prediction Confidence	4.15 (1.40)	4.12 (1.47)
Post-task Prediction Confidence	3.64 (1.59)	3.26 (1.70)

Does confidence track prediction consistency? Based on the self-consistency model of confidence, I hypothesized that participants who expressed the same relatedness beliefs at the pre- and post-task timepoints would be more confident in their final predictions. Thus, I conducted an exploratory 2 (Belief Consistency: Consistent vs. Inconsistent) x 2 (Group: Feedback vs. No Feedback) x 3 (Post-task Belief: Benefit vs. Cost vs. No Difference) ANOVA on participants' post-task prediction confidence. This analysis revealed a main effect of consistency, $F(1, 388) = 70.65$, $MSE = 2.22$, $p < .001$, $\eta_G^2 = .154$, that is, participants who expressed consistent pre- and post-task beliefs were indeed more confident than participants who changed their beliefs. There was also a main effect of feedback: Participants who received feedback were more confident in their final predictions, $F(1, 388) = 4.25$, $MSE = 2.22$, $p = .040$, $\eta_G^2 = .011$. Finally, there was a main effect of relatedness belief direction, $F(2, 388) = 12.89$, $MSE = 2.22$, $p < .001$, $\eta_G^2 = .062$. Pairwise comparisons (Tukey-adjusted) revealed that participants predicting either a benefit [$t(388) = 3.89$, $SE = 0.19$, $p < .001$] or a cost [$t(388) = 4.28$, $SE = 0.21$, $p < .001$] expressed higher confidence than participants predicting no difference. While the three-way interaction was not significant ($p = .301$), there was one significant two-way interaction between consistency and belief direction, $F(2, 388) = 6.11$, $MSE = 2.22$, $p = .002$, $\eta_G^2 = .031$. Follow-up analyses (Tukey-adjusted) suggested that belief consistency was associated with higher confidence for all final belief directions [benefit: $t(388) = 4.83$, $SE = 0.31$, $p < .001$; cost: $t(388) = 5.99$, $SE = 0.36$, $p < .001$; no difference: $t(388) = 3.40$, $SE = 0.22$, $p < .001$], though the effect appeared to be smaller for participants predicting no difference at post-task. Figure 17 shows the relation between final prediction confidence and belief consistency.

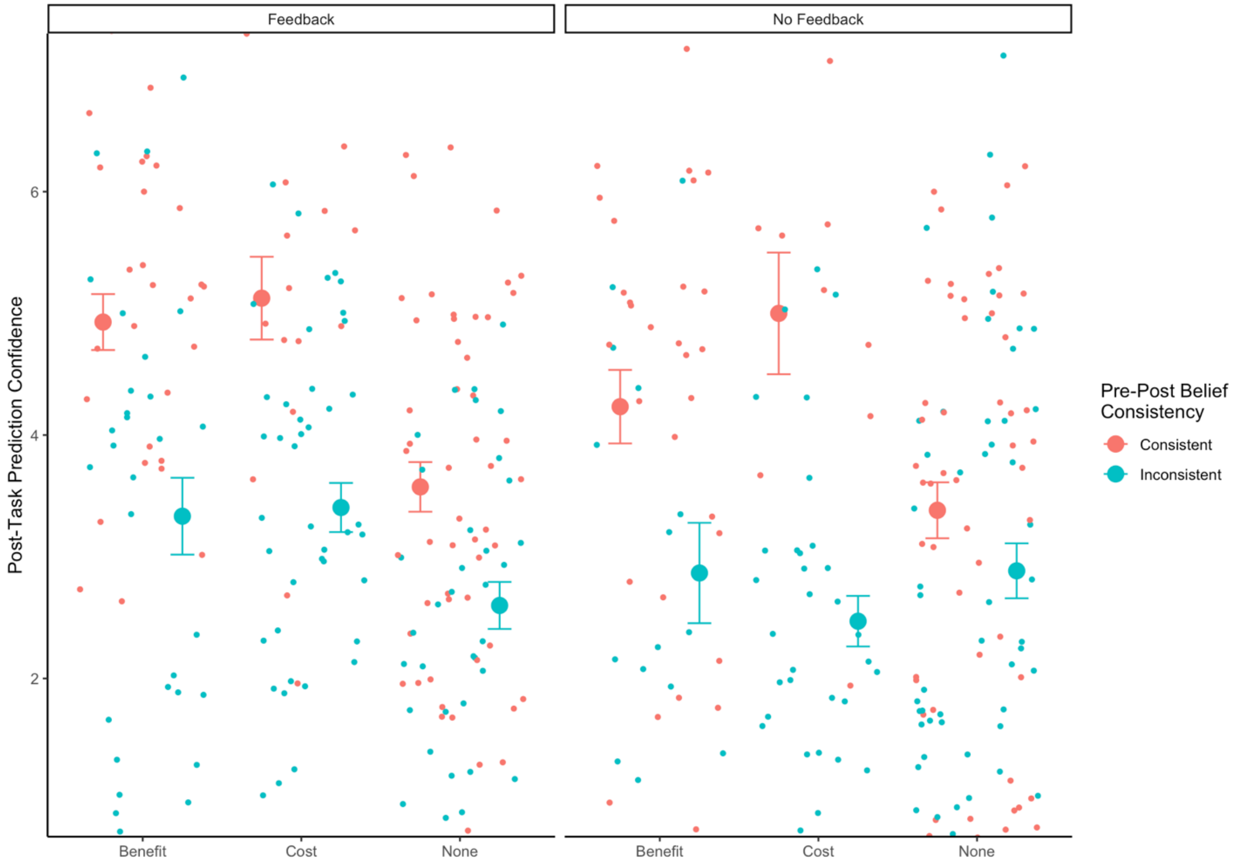


Figure 17. Experiment 11: Participants' post-task prediction confidence as a function of pre-post belief consistency, post-task belief direction, and presence of feedback. Circles in the background represent individual participant confidence.

Does confidence track informational consistency? Based on the self-consistency model of confidence, I further hypothesized that participants who received test experience that was consistent with their pre-task predictions (e.g., predicting a benefit and experiencing a benefit) would have more confidence in their final predictions than would participants who received inconsistent/conflicting test experience (e.g., predicting a benefit but experiencing a cost). Thus, I conducted an exploratory 2 (Initial Belief-Test Experience Consistency: Consistent vs. Inconsistent) x 2 (Group: Feedback vs. No Feedback) x 3 (Pre-task Belief: Benefit vs. Cost vs. No Difference) ANOVA on participants' post-task prediction confidence. This analysis revealed a main effect of consistency, $F(1, 388) = 22.80$, $MSE = 2.48$, $p < .001$, $\eta_G^2 = .055$; participants who experienced the same effect in the test that they had predicted at pre-task were indeed more confident than participants who had test experience that conflicted with their initial belief. The interaction between consistency and pre-task relatedness belief direction was also significant, $F(2, 388) = 5.95$, $MSE = 2.48$, $p = .003$, $\eta_G^2 = .030$, but no other effects were significant (all *p*s > .123). Follow-up analyses (Tukey-adjusted) suggested that belief consistency was associated with higher confidence for participants who had initially predicted a benefit [$t(388) = 4.81$, $SE = 0.26$, $p < .001$] or cost [$t(388) = 3.98$, $SE = 0.46$, $p < .001$], but not for participants who had initially predicted no difference [$t(388) = 0.03$, $SE = 0.37$, $p = .973$]. Figure 18 shows the relation between final prediction confidence and informational consistency.

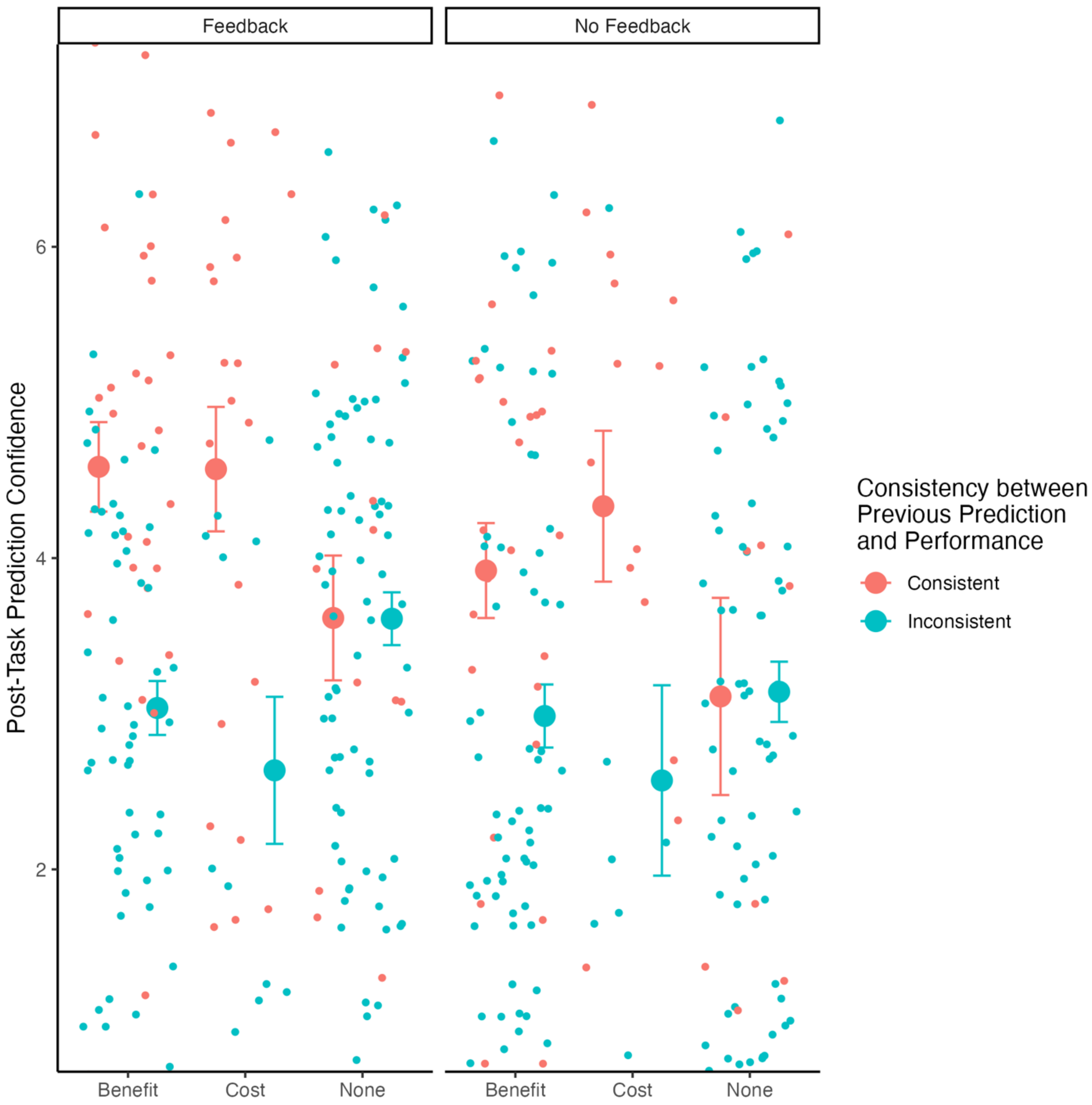


Figure 18. Experiment 11: Participants' post-task prediction confidence as a function of test performance consistency with pre-task belief, pre-task belief direction, and presence of feedback. Circles in background represent individual participant confidence.

Further Exploratory Analyses

Feedback and Postdiction Scores Revisited. Upon inspection of participants' performance vs. postdiction deviation scores, I observed that participants' postdictions tended to be lower than their actual performance (i.e., they underestimated how well they had performed). Thus, I conducted an exploratory 2 (Relatedness: Related vs. Unrelated) x 2 (Group: Feedback vs. No Feedback) ANOVA on participants' performance vs. postdiction deviation scores, preserving the directional difference. The interaction was still not significant, $F(1, 398) = 0.34$, $MSE = 2.73$, $p = .562$, $\eta_G^2 < .001$. The effect of feedback was still significant, $F(1, 398) = 4.51$, $MSE = 5.18$, $p = .034$, $\eta_G^2 = .004$, such that participants who received feedback had deviance scores that were less negative (i.e., participants were more likely to underestimate their performance when they did not receive feedback). However, the effect of relatedness was significant, $F(1, 398) = 27.64$, $MSE = 2.73$, $p < .001$, $\eta_G^2 = .023$, such that participants reported more negative postdictions (relative to their actual performance) for the unrelated condition. In conclusion, participants were more likely to underestimate their performance for the unrelated condition than the related condition, and this effect was not moderated by feedback. One-sample t-tests (Bonferroni-adjusted) suggested that participants' postdiction errors were significantly different from zero for the unrelated conditions ($p < .001$), but not for the related conditions ($p > .190$). Table 38 shows postdiction error and absolute postdiction deviance scores by relatedness and feedback.

Table 38. Experiment 11: Mean (SD) postdiction error (performance – postdiction) and absolute postdiction deviance scores

	Feedback (*N*=205)		**No Feedback (*N*=195)**	
	Related	Unrelated	Related	Unrelated
Postdiction Error	-0.04 (1.78)	-0.59 (1.70)	-0.31 (2.19)	-0.99 (2.25)
Absolute Postdiction Deviance Score	1.18 (1.34)	1.21 (1.32)	1.75 (1.35)	1.85 (1.62)

Experiment 9 Postdiction Scores Revisited. I had not anticipated that participants would underestimate their location memory performance in the unrelated items display. This unexpected result wondering if the same result would be found in Experiment 9, where participants made their postdictions after both study-test blocks but before the post-task prediction. Thus, I conducted an exploratory 2 (Relatedness: Related vs. Unrelated) x 2 (Memory Type: Item vs. Location) ANOVA on participants' postdiction errors in Experiment 9. Critically, the interaction was significant, $F(1, 238) = 13.54$, $MSE = 2.01$, $p < .001$, $\eta_G^2 = .023$. For the item memory task, participants' postdiction errors were more negative for the related condition, $F(1, 121) = 6.38$, $MSE = 1.54$, $p = .013$, $\eta_G^2 = .026$. For the location memory task, participants' postdiction errors were more negative for the related condition, which mirrored the results of Experiment 11, $F(1, 117) = 7.18$, $MSE = 2.49$, $p = .008$, $\eta_G^2 = .023$. One-sample t-tests (Bonferroni-adjusted) suggested that participants' postdiction errors were significantly different from zero for all conditions ($p < .001$) except for unrelated-item memory ($p > .999$). Table 39 shows absolute postdiction deviance scores by relatedness and task type in Experiment 9.

Table 39. Experiment 9: Mean (SD) postdiction error (performance – postdiction)

	Item Memory (*N*=122)		**Location Memory (*N*=118)**	
	Related	Unrelated	Related	Unrelated
Postdiction Error	-0.45 (1.21)	-0.05 (1.26)	-0.63 (1.72)	-1.18 (1.91)

Discussion

In Experiment 11, I once again replicated both the relatedness cost in objective location memory and the relatedness benefit in initial predictions of location memory, as well as the decrease in the predicted relatedness benefit after task experience (Lu et al., 2023; Lu & Risko, under review). As was observed in Experiments 9 and 10, while the predicted relatedness benefit was no longer significant at the post-task timepoint, it did not flip to a cost in the aggregate. Contrary to my initial hypotheses, I did not find that direct test performance feedback moderated the decrease in the predicted benefit of relatedness from pre-task to post-task. While I observed that the proportions of participants in each relatedness belief direction (benefit, cost, no difference) changed from pre-task to post-task, this change was also not moderated by feedback. Feedback did interact with time such that participants' final memory predictions were overall lower than their initial predictions, but they decreased to a lesser extent when participants got feedback than when they did not.

As in the previous two experiments, individual participants shifted their predictions about relatedness in the same direction as their objective test performance; however, contrary to expectations, this effect was not moderated by feedback. I did observe the following effects of feedback: objective relatedness performance was correlated with post-task predictions (consistent with Experiments 1 and 2), but this correlation was stronger for participants receiving feedback. Participants' test performance (related – unrelated) was a strong predictor of final relatedness beliefs (also consistent with Experiments 1 and 2), but this predictive relation was stronger for participants receiving feedback.

As noted above, contrary to my initial hypotheses, I did not find a moderating effect of feedback on the influence of test experience on the change in relatedness beliefs observed from pre-task to post-task, nor did feedback result in a greater decrease in the aggregate predicted benefit of relatedness from pre-task to post-task. However, both of these results are at odds with the finding that feedback did strengthen the contribution of test experience on final beliefs. One possibility is that I was simply unsuccessful in detecting true effects in the belief change regression and aggregate ANOVA. Note that the pre-registered sample size was only determined to be sufficient to detect a difference in the correlation between objective relatedness performance and post-task predictions (assuming that the feedback/no feedback difference would be similar to the item/location difference), and indeed I found that the correlation was stronger for participants receiving feedback. Unfortunately, powering to detect the interaction term in the regression would have taken many hundreds more participants, which was not deemed to be feasible.

Another unexpected finding of Experiment 11 was the lack of a significant correlation between pre-task relatedness predictions and post-task prediction. Note that this correlation was marginal in Experiment 11 ($r = .09$, $p = .062$), which was comparable to Experiment 9 ($r = .18$, $p = .050$), but rather unlike the much stronger correlation observed in Experiment 10 ($r = .47$, $p < .001$). While speculative, perhaps the stronger correlation in Experiment 10 is because participants had to make the post-study judgment, which had the unintended effect of carrying over participants' memory of pre-task judgment (in a 'retrieval practice' sense; Roediger & Butler, 2011), such that they had increased access to their pre-task judgment at post-task. If this hypothesis is correct, then this suggests a tendency for participants to rely upon the previous judgments directly if they are able to retrieve this information. A definitive answer to this awaits exploration in future experiments that manipulate the accessibility of the pre-task prediction information. In any case, the influence of pre-task judgment was not as robust as the influence of the test information on final judgments.

Participants in Experiment 11 were asked to explain both their pre-task and post-task predictions, which offers us additional insight into the different informational bases of each judgment and how their use shifts with experience. For pre-task predictions, participants who predicted a benefit tended to cite the *relatedness improves associations/connections* idea, participants who predicted a cost tended to cite the *relatedness reduces distinctiveness* idea, and participants who predicted no difference tended to cite the *relatedness is irrelevant, memory ability is the most important* idea. This result mirrors previously reported studies (Lu & Risko, under review), supporting the notion that pre-task beliefs about relatedness are a product of different naïve theories that are activated by the task vignette and prompt. For post-task predictions, I observed that task experience was the most commonly cited reasoning across the board (followed by the benefit-via-associations or cost-via-distinctiveness ideas), mirroring the results of Experiment 1. Again, these results suggest that participants considered the test experience as the most important information when constructing their final judgments.

A novel contribution of Experiment 11 was the observation of participants' prediction confidence or *belief credence*. While one might expect that gaining first-hand test information should increase prediction confidence, the opposite occurred: Participants became *less* confident in their post-task predictions compared to their pre-task predictions. I observed that feedback had a moderating effect, such that participants who received feedback reduced their confidence to a lesser extent than participants who did not. These results can be understood within the self-consistency model of subjective confidence (Koriat, 2012; see also Koriat, 2013; 2018; 2024). According to this model, when tasked to answer a particular question, we retrieve a variety of information from memory that is relevant to the question. The level of confidence that we have in our answer is determined by the extent to which our retrieved representations consistently support the chosen answer: If most of the retrieved information points toward the same answer, this is expressed as high confidence; conversely, if the information is contradictory or supports different answers, then confidence will be lower. In the current context, participants' prediction confidence can be taken as a measure of how often they obtain the same answer across their retrieved representations. Thus, at the post-task timepoint, participants have gained additional information (i.e., test experience) that potentially conflicts with the information that their initial pre-task judgments were based on. Participants who received feedback presumably had clearer insight into their test performance, which stabilizes their representation of the test information and consequently mitigates the drop in confidence compared to people who did not get feedback (who presumably experienced additional uncertainty surrounding the test outcomes).

Within the beliefs as constructive judgments framework (Lu & Risko, under review), it is assumed that participants were able to draw on various sources of information (e.g., initial judgments, test experience) when queried for their final beliefs. Applying the self-consistency model of subjective confidence, we should additionally expect participants' final belief credence to track the degree of consistency across these different informational sources. Consistent with this hypothesis, I examined two measures of informational consistency and found that they were both associated with higher final belief confidence. First, people who predicted an effect in same direction at pre-task and post-task (i.e., expressed consistent belief responses) were more confident in their final predictions than were people who changed their prediction direction (i.e., expressed inconsistent belief responses). Second, people who performed in the same direction in the memory test as their pre-task judgment (i.e., received consistent test experience) were more confident in their final predictions than were people who performed in a different direction in the test (i.e., received inconsistent test experience).[12]

[12] Interestingly, this consistency-confidence association was not significant for participants who had initially predicted no effect of relatedness. These participants tend to justify their predictions in terms of their own memory ability and are perhaps less receptive to the influence of experimental conditions (e.g., "my memory is poor regardless of relatedness"; Lu & Risko, under review).

Finally, an unexpected finding was that people underestimated how well they performed in the unrelated items condition (in their postdictions of their performance). In fact, unrelated performance was significantly underestimated (and related performance was not), even when participants got feedback. While the true effect of relatedness at test (related - unrelated) was -0.68 (feedback) and -0.73 (no feedback) in the aggregate, participants' postdictions suggested that they perceived the effect inaccurately: Those who received feedback reported an average effect of -0.31 whereas those who did not receive feedback reported an average effect of only -0.03. Since participants perceived themselves as having done much worse on the unrelated condition than they actually did, their test experience was less effective at 'debiasing' them compared to if they had perceived their performance accurately. This could be an important contributing factor to why participants still did not flip to predicting a relatedness cost at post-task.

General Discussion

In all three experiments, there was a mismatch between participants' initial metamemory predictions and their actual memory performance in the location memory task. At first, upon reading the description of the experimental task, participants predicted that they would remember the locations of more related items than unrelated items. Subsequently, they experienced the location memory task where the aggregate effect was a cost of relatedness, such that participants tended to remember fewer locations for the related items than the unrelated items. When participants' predictions were solicited once more after this experience, there was a decrease in the predicted relatedness benefit, although they did not flip to predicting a relatedness cost in the aggregate. In the following sections, I summarize the empirical results and discuss what they imply for how metamemory beliefs are updated and/or changed with experience.

Test experience as a driver of metamemory belief change

I observed a decrease in participants' predicted relatedness benefit (to non-significance) after they had experience with the location memory test. This decrease was not found when participants made predictions for and experienced an item recall test (Experiment 9). The decrease was also only robust after participants completed the location memory test, and not after the study phase alone (Experiment 10). Critically, the degree of relatedness belief change depended on an individuals' performance in the memory test: Participants who experienced a benefit shifted in the same direction and predicted a greater benefit; participants who experienced a cost shifted in that direction and reduced the predicted benefit. Taken together, these results suggest that participants' test experience was a primary driver of belief change, which was further reflected in their self-reported explanations for their final beliefs. This is consistent with the beliefs as constructive judgments framework (Lu & Risko, under review), where the test information is predicted to contribute significantly to participants' final (post-task) relatedness beliefs because it is recent, highly diagnostic, and accessible.

An unexpected finding was that participants' insight into their test performance was biased. In Experiment 11, I found that people underestimated how well they performed in the unrelated items condition but not the related items condition, even when they received feedback. In Experiment 9, while participants in the location memory group underestimated their performance in both the related and unrelated conditions, they underestimated their performance to a greater extent for the unrelated items display. This "biased performance insight" could be an important contributing factor to why participants did not flip to predicting a relatedness cost at post-task. If people perceive themselves as having done worse in the unrelated condition than they actually did, they are relying on a smaller *subjective* experienced difference of relatedness than their true performance, which leads to them underestimating the cost of relatedness in their final judgment.

I speculate that participants' estimates of performance are biased because they are influenced by their previous judgments (and/or ideas that influenced those judgments). That is, postdictions can be situated within the framework as yet another form of judgment which is inferential in nature and can be influenced if one's prior judgments are taken as informational cues. Thus, since participants had previously accessed the idea that their memory for the unrelated items would be worse, their judgment of their own experience may have been distorted in that same direction. I conducted some exploratory analyses using participants' prior predictions to predict their performance estimates (postdictions) and obtained results that were consistent with this hypothesis. In Experiment 9, after controlling for their true relatedness performance difference, participants' location postdictions were indeed influenced by their pre-task relatedness predictions, $b = 0.21$, $t = 2.08$, $p = .040$. In Experiment 11, after controlling for true relatedness performance[13], participants' location postdictions were not influenced by their pre-task relatedness predictions (*ps* > .519) but were influenced by their post-task relatedness predictions (feedback group: $b = 0.65$, $t = 7.88$, $p < .001$; no feedback group: $b = 0.74$, $t = 13.58$, $p < .001$). It would appear that the main source of distortion in participants' performance estimates comes from them relying on their most recent previous judgment.

[13] The regression model used pre-task predictions, post-task predictions, performance, feedback, and the interaction between performance and feedback to predict postdictions (performance estimates). The interaction between performance and feedback was significant, $b = 1.78$, $t = 3.14$, $p = .002$. The predictive power of true performance was stronger for participants who received feedback, $b = 3.17$, $t = 5.58$, $p < .001$, versus no feedback, $b = 1.10$, $t = 3.28$, $p = .001$.

Informational bases of metamemory belief

Although test performance was consistently a significant contributor to participants' final relatedness beliefs, I also found evidence for the influence of their prior judgments. In Experiment 9, final relatedness predictions were a function of pre-task relatedness predictions even after controlling for objective performance. In Experiment 10, final relatedness predictions were a function of both pre-task and post-study relatedness predictions after controlling for objective performance. In Experiment 11, the contribution of pre-task relatedness predictions did not reach significance in the regression but trended in a consistent direction. However, we are unable to disentangle whether participants were directly retrieving their previous judgments or whether they were retrieving the information that led to those judgments, or some combination of both kinds of information. At present, the framework allows for all these sources of information to contribute to the final expressed belief. I anticipate that future investigations can explore the relative contribution of the various informational sources.

We can gain additional insight into participants' informational cues from their self-reported belief explanations (in Experiments 1 and 3). Previously, Lu and Risko (under review) reported that participants' pre-experience explanations suggested that different individuals were activating distinct naïve theories of memory that led to them to predict different effects of relatedness. People who predicted a benefit tended to invoke different explanatory ideas (i.e., relatedness helps one make associations between items) than did people who predicted a cost (i.e., relatedness reduces item distinctiveness). Experiment 9 adds to this picture in the form of post-task belief explanations: The majority of participants invoked their first-hand experiences to explain their judgments, suggesting that this information became one of the most dominant cues. Nevertheless, participants continued to invoke notions like associations (to explain a benefit) and distinctiveness (to explain a cost). Experiment 11 further allows us to compare participants' pre-task and post-task prediction explanations: Participants seemed to shift from predominantly theory-based explanations (associations vs. distinctiveness) to predominantly experience-based explanations. At post-task, task experience was the most commonly cited reasoning across the board. However, the second most commonly cited reasoning was associations-based (for benefit) and distinctiveness-based (for cost), which are ideas that could be activated at pre-task (or formed during the experiment). These insights paint a picture of final belief reports as judgments that are based upon a variety of informational sources.

Finally, although the focus of this investigation was on beliefs about relatedness (and their subsequent change), there was also observable "belief change" in terms of absolute memory performance. In all three experiments, participants' predictions of location memory decreased from pre-task to post-task, suggesting that this task was more difficult than they had anticipated. In contrast, participants' predictions of item memory (in Experiment 1) increased from pre-task to post-task, suggesting that this task was easier that they had anticipated. Both of these results are consistent with the notion that participants' post-task predictions were to some extent tracking their experiences with the task (since they performed better for item memory and worse for location memory). Participants' predictions of location memory decreased from pre-task to post-study, and post-study to post-test (in Experiment 2), suggesting that they were drawing from both their study experience and their test experience as relevant diagnostic information.

Metamemory belief credence

In the beliefs as constructive judgments framework (Lu & Risko, under review), final (post-task) relatedness beliefs are based upon various informational cues such as their task performance, their previous judgments, and retrieved naïve theories of memory. I found further support for the framework in how participants' prediction confidence or *belief credence* changed after task experience (Experiment 3). In the self-consistency model of subjective confidence (Koriat, 2012; see also Koriat, 2013; 2018; 2024), confidence tracks the extent to which our retrieved representations consistently support a given answer. Within the framework, I posit that participants' confidence in their beliefs (belief credence) is a measure of how often they obtain the same answer across the various informational cues that their belief report judgment is based upon. Thus, participants became less confident in their post-task predictions compared to pre-task predictions because they gained additional information (i.e., test experience) that potentially conflicted with the information that their initial pre-task judgments were based on. Consistent with this hypothesis, people whose test experience conflicted with their initial judgment (e.g., predicted a benefit but experienced a cost) were less confident in their final predictions than were people whose test experience agreed with their initial judgment (e.g., predicted a benefit and experienced a benefit).

The decrease in confidence observed in the current context may have been because participants had been presented with an unusual "metacognitive illusion," such that their test experience tended to conflict with their initial beliefs. Thus, the same strong decrease would not be expected in a context where participants gained experience in the same direction as their initial belief because these informational cues would no longer conflict. I investigated this question with an exploratory analysis that used consistency (between initial belief and test experience) to predict confidence change from pre-task to post-task. Not only was consistency predictive of confidence change, $F(1, 396) = 16.39$, $MSE = 2.48$, $p < .001$, $\eta_G^2 = .040$, but only people whose test experience conflicted with their initial judgment reported a decrease in confidence that was significantly different from zero [$t(287) = 9.55$, $p < .001$, change of -0.91].

People who received consistent test experience did not decrease their confidence [$t(111) = 0.56$, $p = .579$, change of -0.10].

Koriat (2024) has recently proposed that confidence, as a measure of self-consistency, should track the *replicability* of a given response (i.e., how likely a participant would give the same answer when queried at another time). In the current context, I would expect that belief credence should be associated with belief consistency within an individual participant. I found results that supported the hypothesis: People who predicted an effect in the same direction at pre-task and post-task (i.e., reported the same belief at both timepoints) were more confident in their final predictions than were people who changed the direction of their prediction (i.e., reported different beliefs).

A constructive framework for understanding metamemory beliefs

Lu and Risko (under review) recently proposed a new framework for understanding metamemory belief reports as constructive judgments. Here I have shown that the framework can account for empirical patterns in how participants' belief reports change during the course of an experiment. This approach follows in the tradition of inferential cue-utilization approaches in metacognition (Koriat, 1997), as well as research on the bases of belief and belief change in social cognition (Cavanaugh et al., 1998; Crocker et al., 1984; Schwarz & Bohner, 2001; Wyer & Albarracín, 2005). In this closing section, I outline how the framework can help to guide future metacognition research and propose some open questions for investigation.

The beliefs as constructive judgments framework provides a productive lens to understand the role of beliefs in metacognitive judgments. Theoretical accounts of metacognition position beliefs as an important basis for metacognitive judgments such as judgments of learning (dual-basis view: Koriat, 1997; Koriat et al., 2004), or even assume that beliefs are the primary basis of these judgments (analytic processing theory: Mueller & Dunlosky, 2017). I propose that the degree to which a particular belief contributes to a judgment of learning depends on the extent to which each elicitation cues the retrieval of the same kinds of information. Thus, I expect the contribution of any particular expression of belief to vary with the degree of contextual match between the belief's elicitation context and the judgment of learning's elicitation context. For example, we should expect the belief report to be predictive of judgments of learning when the wording of the two prompts is very similar and the judgments occur close in time (e.g., pre-study judgments of learning; Price & Harrison, 2017).

The framework also predicts that the more the items being judged vary in the kinds of informational cues available, the less likely that information associated with the belief for a given cue will be retrieved. For example, Kornell and colleagues (2011) found that labeling items with the number of study repetitions (once vs. twice) did not influence judgments of learning when this factor was manipulated within-subjects in conjunction with actual font size (small vs. large). In their studies, each item was presented with both the font size and repetition information, with font size taking up potentially far more screen space. Presumably, this judgment context was far more likely to cue retrieval of size-related belief information rather than repetition-related belief information. On the other hand, when the authors solicited beliefs about study repetitions in a separate group of participants, this was done via a questionnaire (describing a hypothetical memory experiment) where this was the only cue that was manipulated, which produced a large belief in the effect of repetition.

What happens when a "belief" changes? Some researchers reserve the term "belief" for the estimate of the likelihood that some knowledge is correct or if some event or state of affairs will occur, with the assumption that these are computed online; any retrieved knowledge that goes into this belief computation (e.g., previous judgments, episodes) is termed “belief-relevant information” (Wyer & Albarracín, 2005). In the fragmented model of belief storage (Bendaña & Mandelbaum, 2021), the belief is the same as the collective of information itself – a belief exists in the form of multiple fragments that are all called “beliefs”. What these approaches and the current approach have in common is that any report of a “belief” is a judgment that is constructed from multiple sources of retrieved information. In the current framework, beliefs (operationalized as belief reports) can be changed either through acquiring new information or cues (i.e., fragments), or a change in the weighting of which information is used to construct the belief. In the current investigation, I would argue that participants gained new information during the test which became an important cue when making their final judgment.

When individuals speak of “belief change”, I suspect they do not mean some temporary shift in the information activated at a given moment, but instead some longer lasting and more permanent change in the underlying knowledge structure. These different types of “belief change” could imply different consequences for how long-lasting the observed “change” will be. Previous investigations on mending metacognitive illusions have usually demonstrated changes over a short amount of time (e.g., Koriat & Bjork, 2006a, 2006b; Yan et al., 2016), but it is unknown whether that change persists much longer after the experiment. More broadly, research suggests that attempts to change people’s beliefs do not last long (i.e., interventions have little effect after a delay of over one week; Baesler & Burgoon, 1994). Within the framework, the newly acquired test experience was particularly accessible and influential at the post-task timepoint since it was both highly recent and relevant to the elicitation prompt. However, I expect that access to this information (such as “*I performed poorly in the related condition in this location memory task”*) will diminish over time as the experimental context is left behind. Since pre-task belief reports suggest that people are likely to have more belief fragments that are consistent with a benefit (such as "*Related items are easily associated with each other*"), this leads to the prediction that individuals might go back to "believing" relatedness is beneficial for location memory given sufficient time to "forget" about the specific task experience that they had obtained. More broadly, we can expect any "belief change" that is dependent upon the successful retrieval of a single specific episode to be only transient, with longer-lasting belief change occurring when multiple fragments containing consistent episodes have accumulated.

To conclude, the present framework emphasizes the constructive and cue-dependent nature of belief reports and highlights the importance of context when considering their contribution to metacognitive judgments. Researchers have recently begun to think of beliefs as something that can be empirically studied, leading to an emerging new cognitive science of belief (Porot & Mandelbaum, 2021; Sommer et al., 2022; Van Leeuwen & Lombrozo, 2023). The framework offers a promising direction for future research in metacognition and the broader cognitive science of belief.

Concluding Remarks

This dissertation began by exploring the hypothesis that relatedness would result in a cost to location memory. I then ventured into beliefs about the effect of relatedness and proposed a theoretical framework of metamemory beliefs. The final chapter explored how the framework can be productively applied to understand both metamemory beliefs and belief change. I conclude with a brief summary of the empirical findings and highlight potential avenues for future research.

Relatedness and location memory

Participants tended to perform worse in a location memory task when presented with a display consisting of items from a single category (related) compared to a display consisting of items from different categories (unrelated). A combined analysis across all relevant experiments established a robust cost of relatedness in the location memory task, in contrast to the more typically observed relatedness benefit in item recall.

According to the *semantic interference hypothesis*, while increased semantic similarity across items increases the likelihood of outputting an item in free recall, it also increases interference across these items, which reduces performance when discrimination among the to-be-remembered items is required. Unlike a free recall task, correct performance in a location memory task is contingent upon successful discrimination between items. A strong version of this argument is that semantic similarity-based interference will increase the likelihood that another item is erroneously recovered and recalled in the position of the targeted one (Tse, 2009). In support of this hypothesis, the item location substitution rate (the likelihood of substituting a different item's location for the target location) was indeed higher for the related items display.

An alternative *metacognitive loafing hypothesis* was also considered: that the cost might be driven by participants spending less time and/or effort studying the related items. While the related items display was indeed associated with decreased study time in a self-paced version of the task, study time did not mediate the cost, and there continued to be a significant direct effect of display type on location memory. This does not rule out the possibility that metacognitive loafing might play a role in the cost of semantic relatedness to location memory, as self-paced study time may not fully capture participants' experiences of subjective effort. Nevertheless, study time was not able to explain the relatedness cost that was observed.

These results demonstrate the "double-edged" sword of semantic relatedness in memory: Depending on the task demands, we can observe either memory facilitation based on shared semantic similarity/associations (in the item recall task), or memory impairment based on similarity-based interference (in the location memory task). Participants' pre- and post-task self-reports in Chapters 2 and 3 also illustrate their subjective expectations and experiences of the two effects of relatedness. Before the task, they tended to anticipate memory facilitation based on shared semantic similarity/associations, with more people citing this for item recall. After the task, more participants who had experienced the location memory task were able to describe increased interference stemming from the semantically similar items.

The relatedness halo in memory predictions

While Chapter 1 established an objective cost of relatedness in the location memory task, participants' predictions of memory (as reported in Chapters 2 and 3) did not anticipate this cost. Participants tended to express the general belief that a memory task involving a related list of words would be easier than one involving an unrelated list of words. This effect was found for both item memory (where a benefit is indeed the case) but also location memory (where the opposite is the case). Participants appeared to exhibit a kind of “relatedness halo” in their predictions, predicting that relatedness is beneficial to memory even in a context where it has been demonstrated to be harmful.

Chapter 2 established some key characteristics of this metacognitive bias. When participants were asked to make their predictions for either an item memory task or a location memory task, the predicted relatedness benefit was greater overall for the item memory task than for the location memory task. This effect was driven by more participants predicting a relatedness benefit for item, and more participants predicting a cost for location, rather than them simply predicting a smaller-sized relatedness benefit for location. This seemed to be a difference in belief kind, rather than degree: The task descriptions influenced how likely it was that they would end up predicting that relatedness would result in a benefit, a cost, or no effect.

These three belief kinds were also associated with different kinds of explanations. People who predicted a relatedness benefit tended to reason in terms of memory facilitation based on shared semantic similarity/associations whereas people who predicted a relatedness cost anticipated memory impairment based on similarity-based interference/reduced distinctiveness. Finally, people who predicted no effect of relatedness tended to discount its influence entirely and reasoned instead in terms of their own memory abilities being the limiting factor. When participants were primed with either one of two narratives about memory (associations vs. distinctiveness), this changed the likelihood of what downstream relatedness belief would be expressed. Inducing participants to think about associations in memory led more people to predict a relatedness benefit in the location memory task whereas inducing them to think about distinctiveness in memory led more people to predict a relatedness cost.

Based on these results, I argued that participants' expressed beliefs about relatedness were not static notions but were instead constructed in response to the information that was available in a specific elicitation context. Each task vignette comprises a set of memory cues that can lead to the activation/retrieval of different stored information, and consequently lead to different judgments like a "benefit" or a "cost" of relatedness. While most participants retrieve the idea that related items are more easily connected in memory and end up predicting a benefit, other participants note that the location task demands distinguishing between similar items and use distinctiveness as a criterion, ending up with the prediction of a cost.

As noted in the thesis introduction, memory facilitation based on shared semantic similarity/associations and memory impairment based on reduced distinctiveness are really two sides of the same proverbial coin. However, given that the majority of participants predicted a benefit of relatedness even for the location memory task, people seem to be more likely to retrieve ideas associated with the former than with the latter. When cued with the concept of semantic relatedness in a memory task, participants' default narratives tend to align more closely with association-based ideas (that lead to an expected benefit) rather than with distinctiveness-based ideas (that lead to an expected cost).

The beliefs as constructive judgments framework

The central idea that emerged from Chapter 2 was that the metacognitive "beliefs" reported via memory belief questionnaires can be productively considered as constructive judgments that are responsive to information available in the elicitation context. In Chapter 3, I expanded upon how the *beliefs as constructive judgments* framework can account for how participants' belief reports change in response to experience during the course of an experiment.

Participants' initial metamemory predictions tended to conflict with their actual memory performance in the location memory task: After reading a description of the experiment, they predicted at first that they would remember the locations of more related items than of unrelated items. Subsequently, they experienced the location memory task where the aggregate effect was a cost of relatedness. When participants' predictions were solicited again after this experience, there was a decrease in the predicted relatedness benefit, although they did not reverse to predicting a relatedness cost in the aggregate. This decrease was not found when participants made predictions for and experienced an item recall test and was not found after participants experienced the study phase alone before the location memory test.

The degree of belief change depended on an individuals' performance in the memory test: Participants who experienced a benefit shifted in the same direction and predicted a greater benefit, whereas participants who experienced a cost shifted in that direction and reduced the predicted benefit. Test information, therefore, was a significant driver of changes in individuals' belief reports. This observation is consistent with the beliefs as constructive judgments framework, where the test information is predicted to contribute significantly to participants' final relatedness beliefs because it is recent, highly diagnostic, and accessible. Participants' explanations also tended to cite their experience in the test as the primary basis for their final beliefs, suggesting that this information became one of the most dominant cues. However, their final belief reports were predicted not only by their experiences in the test but also by their previously elicited belief judgments.

The beliefs as constructive judgments framework accounts for these results by assuming that when individuals are probed for their beliefs, their response is constructed based on available information that has been retrieved from memory, such as their previous judgments and other relevant knowledge and episodes. Before experiencing the task, participants are initially more likely to retrieve ideas consistent with memory facilitation based on shared semantic similarity/associations, compared to memory impairment based on similarity-based interference. This may be because experiences pertaining to the former are more common in everyday life, leading to participants having many more relevant instances or episodes to draw upon. However, after experiencing the task first-hand, their performance serves as a highly relevant informational cue for predicting their future performance in a similar task. Participants' explanations for their pre- and post-task judgments were further suggestive of this shift in informational cues: They went from initially citing predominantly theory-based explanations (associations vs. distinctiveness) to predominantly task experience-based explanations, though they continued to espouse theory-based explanations to some extent. These insights paint a picture of final belief reports as judgments that are based upon multiple informational sources.

Altogether, this dissertation makes several novel empirical contributions relevant to our understanding of the influence of relatedness on memory performance and on metamemory beliefs. I have developed a novel theoretical perspective on understanding metamemory beliefs, particularly their expression in response to belief probes. In the following sections, I highlight some additional findings and suggest some areas for future research.

Metamemory belief antecedents and consequents

Where do metamemory beliefs come from, and why – like the relatedness halo – are they sometimes wrong? In Chapter 2, I proposed that participants retrieved distinct naïve theories of memory that led to the prediction of a relatedness benefit or cost (or no difference). One may have wondered where these naïve theories themselves come from. In Chapter 3, I extended this notion of beliefs such that they can be based on various kinds of retrieved information (e.g., previous judgments and relevant episodes). While it is possible that one has a naïve theory or belief in a language-like form ("related items are easier to remember" or "related items are more easily confused"), my framework assumes that for most of our beliefs, this information does not already exist in such a form but instead is constructed online at the time of elicitation based on multiple retrieved instances. In this conception of belief, what distinguishes belief-relevant information from run-of-the-mill memory traces is simply that this information is playing a "belief-like" functional role at a particular moment given particular elicitation cues.

The antecedents of beliefs are therefore an accumulation of our learned experiences combined with our capacity to reason, predict, and act based on this information. The consequences of a metacognitive belief can be observed when we act in a way that is consistent with holding that belief or having that information active. In Experiment 3, participants spent less time studying the related items and more time studying the unrelated items. Their behavior is surely reasonable if we attribute it to them believing in a relatedness benefit (*related items are easier to remember and unrelated items are harder, so I will compensate with study time accordingly*).

The conception of beliefs as constructive judgments suggests that the potential heterogeneity of belief information both across individuals and within an individual will have observable consequences. I would expect that people with different beliefs should tend to act in various ways that are consistent with those beliefs. For example, I would predict that people who believe related items to be easier should study them less than unrelated items, but that people who believe related items to be harder should study them *more* – and I would predict no study time differences in people who think that relatedness does not matter (because they have a fixed notion of their memory ability; Miele & Molden, 2010). Further, we should expect to observe similar consequences when different belief-relevant information is temporarily made salient (i.e., manipulating the antecedents), as in Experiment 8.

Metamemory belief credence

A promising new area for further research concerns participants' confidence in their metacognitive beliefs – their *belief credences*. According to the self-consistency model of subjective confidence (Koriat, 2012), confidence tracks the extent to which our retrieved representations consistently support a given answer. The model further assumes that informational consensus *across individuals* tends to correlate with informational consistency *within* an individual, so people who express the majority opinion tend to be more confident in their answers than people who express a minority opinion. I proposed that this idea can also be applied to the beliefs as constructive judgments framework: One's belief credence can be taken as a measure of consistency across the retrieved informational sources that formed the basis of that belief.

The idea that belief credence tracks one's informational consistency received support from multiple lines of evidence. Initially, more participants predicted a benefit of relatedness compared to a cost, suggesting that most information that can be retrieved tends to point to a benefit rather than a cost. Critically, participants who expressed a relatedness benefit belief (majority opinion) were indeed more confident than participants who expressed a relatedness cost belief (minority opinion). When participants experienced test information that conflicted with their initial beliefs, their belief credence decreased, while this decrease was not observed for participants whose performance was consistent with their initial beliefs. People who expressed different beliefs about the effect of relatedness (at two elicitation timepoints) were also less confident in their predictions than people who consistently expressed the same belief.

To my knowledge, this work represents one of the first investigations of metacognitive belief credence. Given that we may be more likely to act upon high credence beliefs, which can have consequences for metacognitive control decisions (e.g., choosing a particular study strategy; deciding to skip class; spending more time on a particular assignment), belief credence would appear to be an important avenue for further research.

Metacognitive beliefs and judgments of learning

The beliefs as constructive judgments framework emphasizes the constructive and cue-dependent nature of belief reports and highlights the importance of elicitation context. One’s “beliefs” can change depending on seemingly innocuous factors in the elicitation context such as the wording of the question[14], since these will influence retrieval cues and subsequently what information is available and used to construct a belief response.

One prediction that falls out of the framework is that one’s beliefs should have the appearance of “drifting” over time across different spatiotemporal contexts. If an individual is asked for their belief at one time and then another, then the second belief report should resemble the first less and less as the delay between the two judgments increases (i.e., contextual drift; Howard & Kahana, 2002a). At increased delays, the likelihood that an individual will be able to retrieve their previous judgment decreases, and the likelihood that the same kinds of information will be retrieved also decreases. Nonetheless, we should still expect belief consistency within the same individual over time as each judgment will still be based upon sampling from largely the same pool of information.

The framework also provides a productive lens to understand the role of beliefs in metacognitive judgments. Participants’ answers to belief probes and their item-by-item

[14] A large body of work pertains to a similar finding with respect to leading questions in eyewitness memory reports (e.g., Loftus, 1975).

judgments of learning are assumed to result from the same constructive process, except that the latter judgments occur in a different elicitation context where the experience of just having processed a given item is immediate and salient. This leads to the prediction that the degree to which a particular belief contributes to a judgment of learning depends on the extent to which each elicitation cues the retrieval of the same kinds of information. Before considering the contribution of beliefs to judgments of learning, it is important to consider the degree of contextual match between the belief's elicitation context and the judgment of learning's elicitation context.

The emerging cognitive science of belief

As the empirical interest in beliefs has coalesced, a new cognitive science of belief has recently begun to emerge (Porot & Mandelbaum, 2021; Sommer et al., 2022; Van Leeuwen & Lombrozo, 2023). Although the focus of this thesis has been on metacognitive beliefs in a narrow sense (examining just one kind of "belief" about the effect of relatedness on memory), the framework might be productively applied to beliefs in other domains.

In other belief domains (e.g., religious, moral, or political), certain beliefs might be far more entrenched or deeply held, and much less resistant to change. The framework can account for this if we assume that these strong beliefs are represented across many previous judgments, episodes, or instances, such that most of one's retrieved knowledge points in the same belief direction (see also Bendaña & Mandelbaum, 2021).

According to Van Leeuwen and Lombrozo (2023), we ought not to think of a single way to believe, but instead consider different *varieties of believing*. Beliefs might play a variety of different functional roles in cognition, such as epistemic (i.e., factual or truth-tracking) or nonepistemic (identity or in-group affirmation; Metz et al., 2023). While I presume metamemory beliefs to be mainly epistemic in nature, nonepistemic considerations are also possible (e.g., believing that one has a good memory is important to one's identity). Within the framework, these considerations can be conceptualized as different internal goals that serve as distinct retrieval cues or frames.

Conclusion

In this dissertation, I have examined how and why semantic relatedness influences both memory performance and metamemory prediction, and I have introduced a new theoretical framework for understanding metacognitive beliefs. This work represents a novel contribution to the memory and metacognition literature and adds to the emerging cognitive science of belief. In addition, this work has opened a number of doors for future investigation that promise further insights into how we remember and how we think about how we remember.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*(1), 126-131.

Baddeley, A. D. (1966a). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology, 18*(4), 362-365.

Baddeley, A. D. (1966b). The influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology*, *18*(4), 302-309.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47-89). Academic press.

Baesler, E. J., & Burgoon, J. K. (1994). The temporal effects of story and statistical evidence on belief change. *Communication Research, 21*(5), 582-602.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bendaña, J., & Mandelbaum, E. (2021). The fragmentation of belief. In D. Kindermann, C. Borgoni, & A. Onofri (Eds.), *The fragmentation of mind*, Oxford: Oxford University Press.

Bousfield, W. A., Sedgewick, C. H. W., & Cohen, B. H. (1954). Certain Temporal Characteristics of the Recall of Verbal Associates. *The American Journal of Psychology*, *67*(1), 111–118.

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica*, *95*(3), 239-253.

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review*, *14*(1), 107-111.

Cavanaugh, J. C., Feldman, J. M., & Hertzog, C. (1998). Memory beliefs as social cognition: A reconceptualization of what memory questionnaires assess. *Review of General Psychology*, *2*(1), 48-65.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Crocker, J., Fiske, S. T., & Taylor, S. E. (1984). Schematic bases of belief change. In J. R. Eiser (Ed.), *Attitudinal judgment* (pp. 197-226). Springer New York.

Dunlosky, J., Mueller, M. L., & Tauber, S. K. (2015). In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger, III (Eds.), *Remembering: Attributions, processes, and control in human memory* (pp. 46-63). Psychology Press.

Epstein, M. L., Phillips, W. D., & Johnson, S. J. (1975). Recall of related and unrelated word pairs as a function of processing level. *Journal of Experimental Psychology: Human Learning and Memory*, 1(2), 149–152.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London, UK: SAGE.

Gardiner, J. M., Craik, F. I., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 778-783.

Guérard, K., & Saint-Aubin, J. (2012). Assessing the effect of lexical variables in backward recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 312.

Guerin, S. A., & Miller, M. B. (2008). Semantic organization of study materials has opposite effects on recognition and recall. *Psychonomic Bulletin & Review*, *15*(2), 302–308.

Hirtle, S. C., & Mascolo, M. F. (1986). Effect of semantic clustering on the memory of spatial locations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(2), 182–189.

Hourihan, K. L., & Tullis, J. G. (2015). When will bigger be (recalled) better? The influence of category size on JOLs depends on test format. *Memory & Cognition*, *43*(6), 910–921.

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of mathematical psychology*, *46*(3), 269-299.

Howard, M. W., & Kahana, M. J. (2002b). When Does Semantic Similarity Help Episodic Retrieval? *Journal of Memory and Language*, *46*(1), 85–98.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67*(3), 247-257.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576-90.

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology, 86*(5), 680.

Hund, A. M., & Plumert, J. M. (2003). Does Information About What Things Are Influence Children's Memory for Where Things Are? *Developmental Psychology*, *39*(6), 939–948.

Hunt, R. R. (2013). Precision in Memory Through Distinctive Processing. *Current Directions in Psychological Science*, *22*(1), 10–15.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 497–514.

Hunt, R. R., & McDaniel, M. A. (1993). The Enigma of Organization and Distinctiveness. *Journal of Memory and Language*, *32*(4), 421–445.

Ishiguro, S., & Saito, S. (2021). The detrimental effect of semantic similarity in short-term memory tasks: A meta-regression approach. *Psychonomic Bulletin & Review*, *28*(2), 384-408.

Jackson, E. (2022). On the independence of belief and credence. *Philosophical Issues*, *32*(1), 9-31.

Kahana, M. J. (2020). Computational Models of Memory Search. *Annual Review of Psychology*, *71*(1), 107–138.

Kahana, M. J., Diamond, N. B., & Aka, A. (2022). *Laws of Human Memory* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/aczu9

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM Model of False Recall. *Psychological Review*, *114*(4), 954–993.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187–194.

Koriat, A., & Bjork, R. A. (2006a). Illusions of competence during study can be remedied by manipulations that enhance learners’ sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959-972.

Koriat, A., & Bjork, R. A. (2006b). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133–1145.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: the role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*(4), 643-656.

Koriat, A. (2013). Confidence in personal preferences. *Journal of Behavioral Decision Making*, *26*(3), 247-259.

Koriat, A. (2024). Subjective confidence as a monitor of the replicability of the response. *Perspectives on Psychological Science*, 17456916231224387.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80-113.

Koriat, A. (2018). When reality is out of focus: Can people tell whether their beliefs and judgments are correct or wrong? *Journal of Experimental Psychology: General*, *147*(5), 613-631.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*(6), 787-794.

Kowialiewski, B., Krasnoff, J., Mizrak, E., & Oberauer, K. (2023). Verbal working memory encodes phonological and semantic information differently. *Cognition*, 233, 105364.

Lakens, D., & Caldwell, A. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592095150.

Lewis, M. Q. (1971). Categorized lists and cued recall. *Journal of Experimental Psychology*, *87*(1), 129–131.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*(4), 560-572.

Lu, X., Bhandari, B., & Risko, E. F. (under review). How do metamemory beliefs change with experience? A constructive, cue-dependent framework.

Lu, X., Kelly, M. O., & Risko, E. F. (2022). The gist of it: Offloading memory does not reduce the benefit of list categorisation. *Memory*, *30*(4), 396–411.

Lu, X., Penney, T. B., & Kang, S. H. K. (2021). Category similarity affects study choices in self-regulated learning. *Memory & Cognition, 49*(1), 67–82.

Lu, X., Zhu, M. J. H., & Risko, E. F. (2023). Semantic relatedness can impair memory for item locations. *Psychological Research*. Advance online publication.

Lu, X., Zhu, M. J. H., & Risko, E. F. (2024). Semantic partitioning facilitates memory for object location through category-partition cueing. *Memory*. Advance online publication.

Lu, X., & Risko, E. F. (under review). The relatedness halo in predictions of learning.

Lüdecke, D. (2022). *sjPlot: Data Visualization for Statistics in Social Science.* (2.8.12) [Computer software]. https://cran.r-project.org/package=sjPlot

Mandler, G. (1967). Organization and Memory. In K. W. Spence & J. T. Spence (Eds.), *Psychology of Learning and Motivation* (Vol. 1, pp. 327–372). Academic Press.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Matvey, G., Dunlosky, J., & Schwartz, B. (2006). The effects of categorical relatedness on judgements of learning (JOLs). *Memory*, *14*(2), 253–261.

Metz, S. E., Liquin, E. G., & Lombrozo, T. (2023). Distinct Profiles for Beliefs About Religion Versus Science. *Cognitive Science*, *47*(11), e13370.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition, 39*(3), 462–476.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, 20, 378-384.

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2016). The effect of identical word pairs on people's metamemory judgments: What are the contributions of processing fluency and beliefs about memory?. *Quarterly Journal of Experimental Psychology*, 69(4), 781-799.

Neale, K., & Tehan, G. (2007). Age and redintegration in immediate memory and their relationship to task difficulty. *Memory & Cognition*, 35, 1940-1953.

Neath, I., Saint-Aubin, J., & Surprenant, A. M. (2022). Semantic relatedness effects in serial recall but not in serial reconstruction of order. *Experimental Psychology*, *69*(4), 196.

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & Cognition*, *41*(6), 797–819.

Poirier, M., & Saint-Aubin, J. (1995). Memory for Related and Unrelated Words: Further Evidence on the Influence of Semantic Factors in Immediate Serial Recall. *The Quarterly Journal of Experimental Psychology Section A*, *48*(2), 384–404.

Porot, N., & Mandelbaum, E. (2021). The science of belief: A progress report. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(2), e1539.

Postma, A., & De Haan, E. H. F. (1996). What Was Where? Memory for Object Locations. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 178–199.

Postma, A., Kessels, R. P. C., & van Asselen, M. (2008). How the brain remembers and forgets where things are: The neurocognition of object–location memory. *Neuroscience & Biobehavioral Reviews*, *32*(8), 1339–1345.

Puff, C. R. (1970). Role of clustering in free recall. *Journal of Experimental Psychology*, *86*(3), 384–386.

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1–2), 103–121.

Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, *175*, 2353-2372.

R Core Team. (2019). *R: A language and environment for statistical computing* (3.6.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting Clustering From Semantic Structure. *Psychological Science*, *4*(1), 28–34.

Saint-Aubin, J., Ouellette, D., & Poirier, M. (2005). Semantic similarity and immediate serial recall: Is there an effect on all trials. *Psychonomic Bulletin & Review*, *12*(1), 171–177.

Saint-Aubin, J., & Poirier, M. (1999). Semantic similarity and immediate serial recall: Is there a detrimental effect on order information? *The Quarterly Journal of Experimental Psychology: Section A*, 52(2), 367-394.

Schwarz, N., & Bohner, G. (2001). The construction of attitudes. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intrapersonal processes* (Vol. 1, pp. 436–457). Oxford, England: Blackwell.

Siegel, A. L. M., & Castel, A. D. (2018). Memory for Important Item-Location Associations in Younger and Older Adults. *Psychology and Aging*, *33*(1), 30–45.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. (2023). *afex: Analysis of Factorial Experiments.* (1.3-0). [Computer software]. https://cran.r-project.org/package=afex

Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(5), 1236–1242.

Sommer, J., Musolino, J., & Hemmer, P. (2022). Toward a cognitive science of belief. In J. Musolino, J. Sommer, & P. Hemmer, (Eds.), *The cognitive science of belief: A multidisciplinary approach* (pp. 2–23). Cambridge University Press.

Surprenant, A. M., Neath, I., & Brown, G. D. A. (2006). Modeling age-related differences in immediate memory using SIMPLE. *Journal of Memory and Language*, *55*(4), 572–586.

Tauber, S. K., Dunlosky, J., Rawson, R. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review, 20*, 356–363.

Tompary, A., & Thompson-Schill, S. L. (2021). Semantic influences on episodic memory distortions. *Journal of Experimental Psychology. General*, *150*(9), 1800–1824.

Tse, C.-S. (2009). The role of associative strength in the semantic relatedness effect on immediate serial recall. *Memory*, *17*(8), 874–891.

Tse, C.-S., Li, Y., & Altarriba, J. (2011). The effect of semantic relatedness on immediate serial recall and serial recognition. *Quarterly Journal of Experimental Psychology*, *64*(12), 2425–2437.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381–391.

Underwood, B. J. (1957). Interference and forgetting. *Psychological review*, *64*(1), 49.

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, *43*, 647-658.

Van Leeuwen, N., & Lombrozo, T. (2023). The puzzle of belief. *Cognitive science*, *47*(2), e13245.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335.

Vuorre, M. (2016). *bmlm: Bayesian Multilevel Mediation* (1.3.12) [Computer software]. https://cran.r-project.org/package=bmlm

Wyer, R. S., Jr., & Albarracín, D. (2005). Belief formation, organization, and change: Cognitive and motivational influences. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change* (pp. 273–322). Mahwah, NJ: Erlbaum.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General, 145*(7), 918–933.

Yang, C., Huang, T. S.-T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language*, *99*, 99–110.

Yang, C., Yu, R., Hu, X., Luo, L., Huang, T. S.-T., & Shanks, D. R. (2021). How to assess the contributions of processing fluency and beliefs to the formation of judgments of learning: Methods and pitfalls. *Metacognition and Learning*, *16*(2), 319–343.

Appendices

Appendix A

Word lists used as stimuli

Experiments 1a and 1b

wearable	**tools**	**toiletries**	**office**	**kitchen**	**instruments**
sneakers	cutters	hairbrush	eraser	spatula	saxophone
necklace	pliers	deodorant	binder	muffin	banjo
sweater	screwdriver	toothpaste	clipboard	kettle	clarinet
sunglasses	scissors	lotion	ruler	sponge	harp
socks	wrench	floss	pencil	mug	violin
gloves	axe	comb	calculator	toaster	flute
jeans	drill	razor	notebook	spoon	trumpet
shorts	hammer	towel	envelope	tray	drum
shirt	pocketknife	mouthwash	folder	whisk	piano
belt	bolt	soap	printer	pan	bass

Experiment 2

animal
bear
cat
cow
deer
dog
elephant
horse
lion
pig
tiger

building
apartment
cabin
condo
dorm
house
hut
mansion
shack
tent
trailer

carpentry
chisel
drill
hammer
nail
ruler
sander
saw
screw
screwdriver
wrench

clothing
hat
jacket
pants
shirt
shoes
shorts
skirt
socks
sweater
underwear

fruit
apple
banana
grape
kiwi
orange
peach
pear
pineapple
strawberry
watermelon

reading
article
book
journal
letter
magazine
newspaper
novel
pamphlet
textbook
website

furniture
bed
chair
couch
desk
dresser
lamp
loveseat
sofa
stool
table

kitchen
bowl
fork
knife
ladle
pan
plate
pot
spatula
spoon
whisk

time
century
day
decade
hour
millisecond
minute
month
second
week
year

flavoring
butter
garlic
ketchup
mustard
onions
pepper
salt
spices
sugar
vanilla

relative
aunt
brother
cousin
father
grandfather
grandmother
mother
niece
sister
uncle

Experiment 3

wearables	**tools**	**toiletries**	**office**
sneakers	cutters	hairbrush	eraser
necklace	pliers	deodorant	binder
sweater	screwdriver	toothpaste	clipboard
sunglasses	scissors	lotion	ruler
socks	wrench	floss	pencil
gloves	axe	comb	calculator
jeans	drill	razor	notebook
shorts	hammer	towel	envelope
shirt	pocketknife	mouthwash	folder
belt	bolt	soap	printer

kitchen	**instruments**	**toys**	**sports**
spatula	saxophone	chessboard	basketball
plate	banjo	dice	bicycle
kettle	clarinet	doll	frisbee
tongs	harp	kite	helmet
mug	violin	puzzle	jumprope
toaster	flute	scrabble	shuttlecock
spoon	trumpet	slinky	skateboard
tray	drum	uno	skis
whisk	piano	lego	trampoline
pan	bass	checkers	volleyball

furniture	**pantry**	**cleaning**
table	pasta	mop
couch	rice	vacuum
lamp	milk	detergent
nightstand	sugar	rag
shelf	flour	duster
cabinet	bread	disinfectant
sofa	cereal	dustpan
stool	salt	broom
carpet	pepper	bleach
armchair	chips	sponge

Appendix B

Narrative Primes used in Experiment 8

Associations Prime:

Research has consistently shown that the ability to form associations is very important for memory. According to the associative network theory, memory is organized as a network of interconnected nodes, where each node represents a concept or piece of information that can be linked to other nodes. Memory is therefore inherently associative: we learn new information by linking it together with existing information in memory, creating a web of connections.

Forming and retrieving associations is the key to how memory works. For example, if you need to memorize a list of tasks to do, consciously creating connections between the items will help you to remember them later. Understanding the ability to create associations and using this knowledge can be beneficial in various aspects of life, from improving study habits to enhancing everyday memory. When we consciously make connections between new pieces of information and integrate them into our existing knowledge, we enhance memory and our ability to remember information more effectively.

References:

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261-295.

Anderson, J. R., & Bower, G. H. (2014). *Human associative memory*. Psychology Press.

Distinctiveness Prime:

Research has consistently shown that the ability to distinguish between different pieces of information is very important for memory. According to the interference theory of memory, we forget things—or have difficulty remembering things—because old and new information interfere with each other in memory. We can overcome this by taking advantage of distinctiveness—because distinctive information is remembered better than information that is not distinctive. Distinctiveness can be perceptual, such as a red-coloured item in the context of blue-coloured items, or it can be conceptual, such as a number embedded in a row of letters.

Overcoming interference via distinctiveness is the key to how memory works. For example, if you take the same route to work every day, unless something new or unique happens, you may be unable to remember the specific details of each trip. Understanding distinctiveness and using this knowledge can be beneficial in various aspects of life, from improving study habits to enhancing everyday memory. When we consciously make pieces of information distinctive by focusing on the differences between them, we enhance memory and our ability to remember information more effectively.

References:

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558-578.

Hunt R. R. (2006). The concept of distinctiveness in memory research. In Hunt R. R. & Worthen J. B. (Eds.), *Distinctiveness and memory* (pp. 3–25). New York: Oxford University Press.

Appendix C

Detailed objective memory performance analyses for Experiment 11

In Chapter 3, I reported the subject-level location accuracy means for Experiment 11. Since one of the pre-registered goals of this experiment was to determine if the location memory results from Chapter 1 could be replicated, here I report the same trial-level performance measures from that chapter: (1) location accuracy, a binary variable indicating whether participants selected the correct location; (2) Euclidean distance, a continuous variable measuring the distance from participants' chosen location to the correct location; (3) item substitution rate, defined as a binary variable indicating whether participants were more likely to select the location of another item (instead of a blank space) when making an error.

I first conducted analyses on the final set of participants after all exclusions (*N* = 400), and then repeated them on data that included participants who had been excluded only for failing the final prediction comprehension checks (*N* = 145 exclusions, total *N* = 545). Table 40 shows the three objective memory measures obtained from the post-exclusions dataset; Table 41 shows the same measures from the pre-exclusions dataset.

Table 40. Experiment 11: Mean (SD) objective memory measures on post-exclusions data

	Related	**Unrelated**
Location Accuracy	.40 (.23)	.47 (.24)
Euclidean Distance	1.37 (0.68)	1.15 (0.65)
Item Substitution Rate	.44 (.23)	.41 (.26)

Table 41. Experiment 11: Mean (SD) objective memory measures on pre-exclusions data

	Related	**Unrelated**
Location Accuracy	.39 (.24)	.45 (.25)
Euclidean Distance	1.38 (0.68)	1.20 (0.68)
Item Substitution Rate	.44 (.23)	.41 (.25)

Location Accuracy. A mixed-effects logistic regression revealed a significant main effect of relatedness, such that location memory accuracy was lower for the related items than unrelated, b = -0.33, 95% CI [-0.45, -0.21], z = 5.29, p < .001. A similar result was obtained with the pre-exclusions data, b = -0.25, 95% CI [-0.36, -0.14], z = 4.54, p < .001. Participants were less likely to select the correct item location when the items were related.

Euclidean Distance. A mixed-effects linear regression revealed a significant main effect of relatedness, such that the mean Euclidean distance between selected and target locations was greater for the related items than unrelated, b = 0.21, 95% CI [0.15, 0.26], t = 6.95, p < .001. A similar result was obtained with the pre-exclusions data, b = 0.16, 95% CI [0.11, 0.21], t = 6.35, p < .001. Participants' answers tended to be locations that were further away from the target location when the items were related.

Item Substitution Rate. A mixed-effects logistic regression revealed a significant main effect of relatedness: participants were more likely to select another item on incorrect trials for the related items display, b = 0.15, 95% CI [0.03, 0.27], z = 2.44, p = .015. A similar result was obtained with the pre-exclusions data, b = 0.14, 95% CI [0.04, 0.25], z = 2.82, p = .003.

In sum, participants performed worse in the location memory task when the items were related. This result was obtained using both a binary measure of accuracy and a continuous measure of distance. In support of the interference hypothesis, participants were also more likely to substitute a related item location when making an error (e.g., choosing the square that had contained *kettle* instead of *mug*), suggesting that the related items were more easily confused with each other (increased inter-item interference) compared to the unrelated items.