Verbal Working Memory Encodes Phonological and Semantic Information Differently

Kowialiewski, B.[1,2], Krasnoff, J.[1], Mizrak, E.[1,3], & Oberauer, K.[1]

[1] Department of Psychology, University of Zurich, Switzerland

[2] University of Liège, Liège, Belgium

[3] Department of Psychology, University of Sheffield, United Kingdom

Correspondence concerning this article should be addressed to Benjamin Kowialiewski, Department of Psychology, Cognitive Psychology Unit, University of Zurich, Binzmühlestrasse 14/22, 8050 Zurich, Switzerland. E-mails : benjamin.kowialiewski@uzh.ch; bkowialiewski@uliege.be

Open Science statement:

All the data and codes have been made available on the Open Science Framework: https://osf.io/tpsg2/

Abstract

Working memory (WM) is often tested through immediate serial recall of word lists. Performance in such tasks is negatively influenced by phonological similarity: People more often get the order of words wrong when they are phonologically similar to each other (e.g., cat, fat, mat). This phonological-similarity effect shows that phonology plays an important role for the representation of serial order in these tasks. By contrast, semantic similarity usually does not impact performance negatively. To resolve and understand this discrepancy, we tested the effects of phonological and semantic similarity for the retention of positional information in WM. Across six experiments (all Ns = 60 young adults), we manipulated between-item semantic and phonological similarity in tasks requiring participants to form and maintain new item-context bindings in WM. Participants were asked to retrieve items from their context, or the contexts from their item. For both retrieval directions, phonological similarity impaired WM for item-context bindings across all experiments. Semantic similarity did not. These results demonstrate that WM encodes phonological and semantic information differently. We propose a WM model accounting for semantic-similarity effects in WM, in which semantic knowledge supports WM through activated long-term memory.

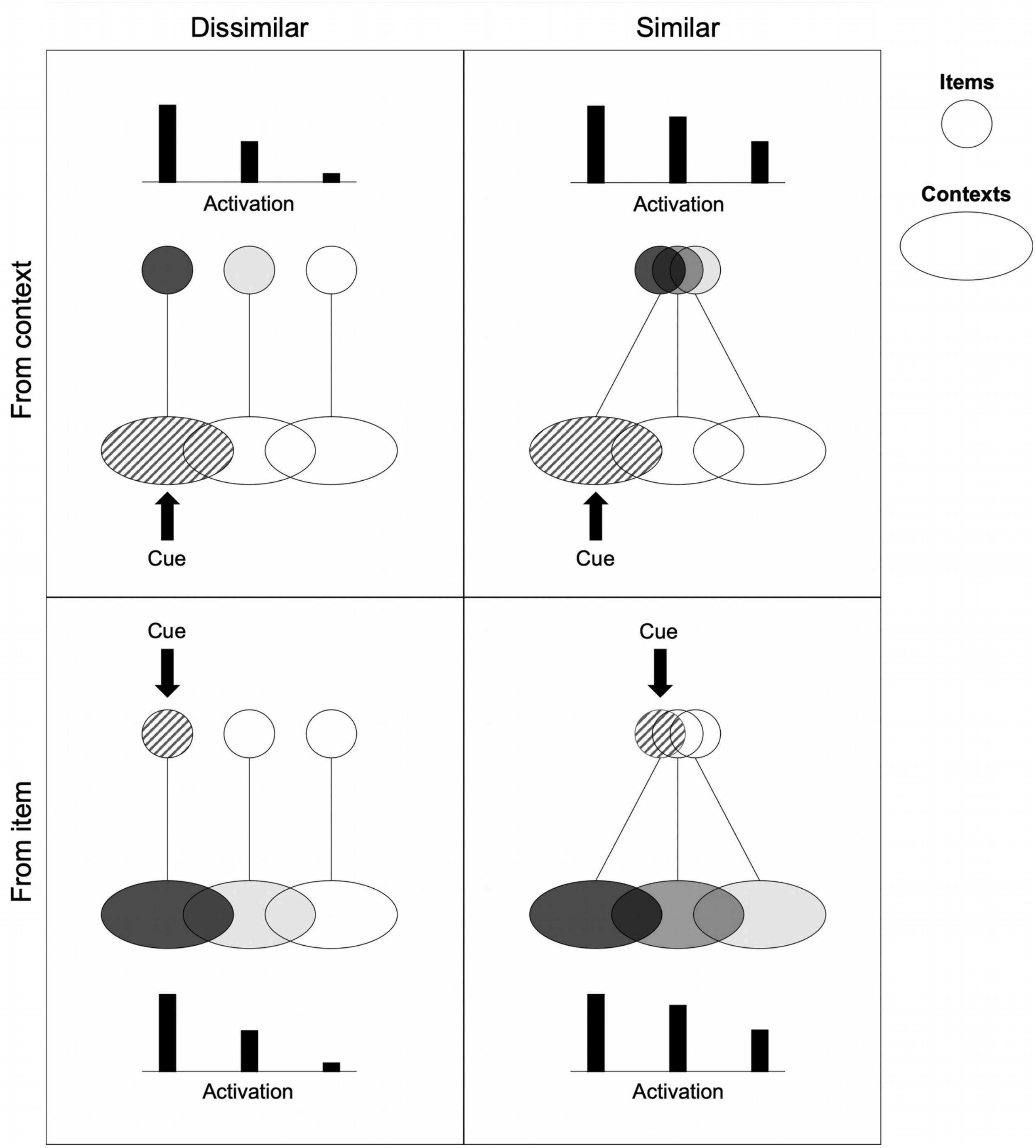
Keywords: Working Memory; Binding; Phonology; Semantic

Working memory (WM) is a core function of the cognitive system responsible for holding information briefly available for further processing. It has long been shown that the phonological similarity between items in a to-be-remembered list induces confusion errors (Baddeley, 1966). When participants study lists such as “rat, fat, mat” and are asked to recall them in serial order, they confuse the order of these words more often compared to lists such as “wall, dig, bend”. Semantically similar lists such as “leopard, cheetah, lion”, by contrast, do not reliably lead to such confusion errors compared to semantically dissimilar lists, such as “sky, pen, pillow” (Saint-Aubin & Poirier, 1999b). In this work, we comprehensively tested the boundary conditions in which semantic similarity could induce confusion errors. Based on our results, we arrived at the conclusion that semantic and phonological information play different roles in the short-term maintenance of serial/positional information.

This study is motivated by models postulating an item’s position in WM is maintained through item-context binding, as implemented in many computational models of serial recall (Burgess & Hitch, 1999, 2006; Henson, 1998; Lewandowsky & Farrell, 2008; Oberauer et al., 2012; Oberauer & Lewandowsky, 2011). In these models, serial position is temporarily maintained by binding items – such as words – to contexts – such as a word’s serial position in a list (e.g., binding the word “wall” to “Position 1”). We illustrate this assumption in **Figure 1**. Suppose the to-be-remembered sequence is “wall, dig, bend”. If asked to recall the item that was presented in the third position, one can re-activate the context of third position and use it as cue to retrieve the word “bend” that is bound to it (**Figure 1**, top). Likewise, if asked to recall where “dig” was presented, one can retrieve “Position 2” (**Figure 1**, bottom). The generic associative model in **Figure 1** allows this flexibility: Retrieving an item when cued with a context/position, but also retrieving a context/position when cued with an item. It is this item-context binding that we assume is responsible for maintaining the item's serial order in a list. This assumption is supported by modelling work that has identified item-context bindings as an essential component of working memory for lists (Farrell & Lewandowsky, 2004) as well as for visual-spatial arrays (Oberauer & Lin, 2017; Schneegans & Bays, 2017).

Figure 1

Illustration of the Binding Process and its Interaction with Similarity.

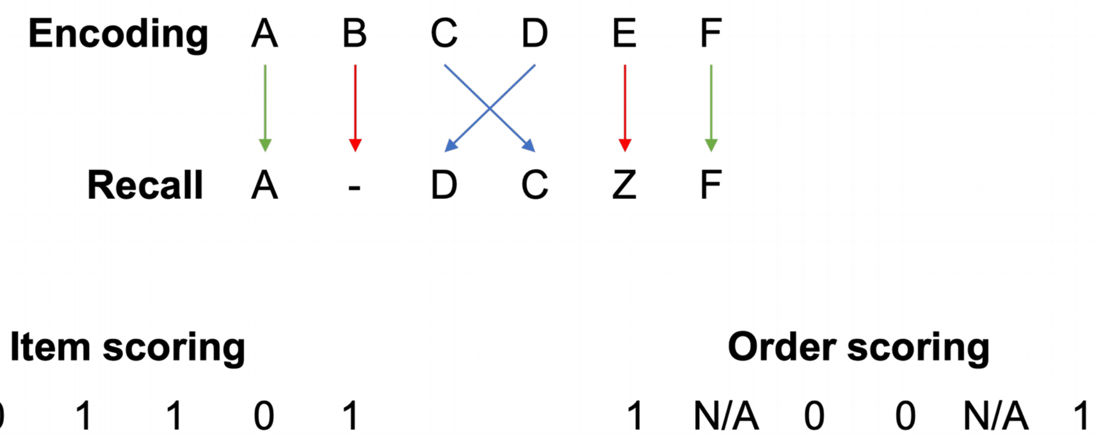


Note. *Through temporary bindings in WM, items (here depicted as circles) can be retrieved from their context (upper panels), and contexts (here depicted as ellipses) can be retrieved from their item (lower panels). When items are dissimilar (left panels), they are sufficiently distinct to allow the original item or context to be retrieved in most cases. When* *inter-item similarity increases (right panel), the competition between alternative WM representations increases, increasing the probability that a confusion error occurs.*

It has long been established that the item-context bindings are subject to confusion errors (Henson, 1998). In the serial-recall literature, these errors are typically referred to as *order errors*, which is a specific type of confusion error appearing in tasks where people need to recall the items in their serial position of a list. When recalling lists of words in their serial order, people often recall the correct words from the list but in wrong list positions. For instance, when trying to recall the sequence “wall, dig, bend” people sometimes retrieve “wall, bend, dig” instead. These order errors are more likely to occur between items sharing adjacent vs. distant serial positions in the list (i.e., the *locality constrain*, see Henson 1998), suggesting some degree of overlap between adjacent positional representations (i.e., the overlapping ellipses in **Figure 2**). These confusion errors must be distinguished from item errors, which is the failure to recall an item at all. One might not be able to recall “bend” and either respond with another word that did not exist in the list or leave their response empty. These errors and the way to compute them are illustrated in **Figure 2**. Studies have shown a dissociation between confusion and item errors. Confusion errors are more affected by dual-task interference than item errors are (Gorin et al., 2016; Henson et al., 2003), and they are associated with different neural regions (Kalm & Norris, 2014; Majerus et al., 2010). Confusion errors are particularly diagnostic to understand what kind of representations is bound to context, because they reflect failures of distinctly binding each item to its context. We will therefore focus on these errors when examining the role of phonological and semantic representations for the item-context binding process. The role of item memory will be considered in the Discussion.

Figure 2

Scoring Procedure Typically Used to Assess Item and Order Memory



Note. *When measuring participants' ability to recall items, the total number of items recalled is computed, divided by the number of memoranda. In this example, four items (A, C, D and F) out of six (A, B, C, D, E, and F) have been recalled, leading to an item score of 4/6 = 0.667. When measuring participants' ability to recall the order of a sequence, the number of items recalled in their correct position is computed, divided by the total number of items recalled regardless of their position. In this example, only two items have been recalled in their correct position (A and F), out of four items in total (A, C, D, and F), leading to an order score of 2/4 = 0.5. When computing the order score, items not recalled at all are scored as missing values. As these items are not recalled at all, they are not informative regarding participants' ability to recall the items in their order. In this example, items B and E have not been recalled at all, they are therefore scored as missing values. In this way, the order score is independent of the item score: A person can have any order score between 0 and 1 regardless of how many items they recalled (when no* *item was recalled, the order score is not defined). See also the Methods section for a detailed description of how these scores were obtained in our experiments.*

The similarity between to-be-remembered information impacts confusion errors, with similar information being more confusable than dissimilar one. The best studied example of this phenomenon is the phonological similarity effect (Baddeley, 1966), in which phonologically similar list items are confused more often than phonologically dissimilar items. This similarity effect is of critical importance for our understanding of WM. It shows that the phonological representation of items is bound to positional contexts. This impact of between-item similarity has been observed across multiple domains, such as the auditory (Visscher et al., 2007; Williamson et al., 2010), and visual (Guitard & Cowan, 2020; Jalbert et al., 2008; Logie et al., 2016; Saito et al., 2008) ones. Therefore, the increased confusability induced by similarity appears to reflect a general property of WM. Confusion errors in WM can be more generally attributed to a discriminability problem. Whatever representation is used during item-context binding, this representational format is subject to confusion errors, especially when the to-be-remembered information becomes difficult to discriminate (i.e., as similarity increases).

Similarity-based Confusions and the Direction of Retrieval

The similarity between items can cause confusion errors in two different ways. The first one occurs when items need to be retrieved from their context, such as retrieving “wall” from “Position 1”. This is the best studied case of similarity-based confusion in the WM literature, in which the so-called phonological similarity effect occurs. The second case is rarely studied in the WM literature and involves retrieving a position/context from the item it was bound to, such as retrieving "Position 2" when presented with "dig". In this section, we explain more thoroughly each type of retrieval direction and the way it is affected by similarity.

When items need to be retrieved from their context, similarity increases confusions because the retrieved WM trace is ambiguous compared to other items (see **Figure 1**, upper panels). For instance, in serial recall, participants must reproduce the items in order. In serial-recall models, this is accomplished by re-activating the positions one by one in forward order and using each position as a cue to retrieve the item bound to it (e.g., "Position 1" is used as cue to retrieve "rat"). This initially leads to the retrieval of a partially degraded WM trace of the item. To produce a legitimate response (e.g., a word), the degraded WM traces must be disambiguated by comparing them to a set of response candidates (Schweickert, 1993). Between-item similarity increases confusion errors during this disambiguation stage. For instance, given the item "rat" and its degraded trace "_at", it is more likely to select "fat" than "dig".

The opposite direction of retrieval is to provide an item and ask to retrieve the position associated to that item. This direction of retrieval – rarely tested in the WM literature – provides a new way for testing a prediction from the idea of item-context bindings, as shown in **Figure 1**. For this direction of retrieval, higher between-item similarity is predicted to increase the probability of confusion errors because the cue itself (i.e., the item) is similar to other cues (i.e., other items in the list) (Mensink & Raaijmakers, 1988; Osgood, 1949; Watkins & Watkins, 1976), and therefore more ambiguous. We will refer to this phenomenon as the *cue-similarity principle*. During the binding process, all features of an item are bound to the item's context. Similar items have overlapping features. When the item features are activated by the item cue, because of the overlapping features, this activates other items' contexts as well as the target item's context. The activation of multiple contexts by the same item cue increases retrieval competition, and hence, the probability of choosing the non-target context. For instance, when presented with the item “rat”, and the next list word was “fat”, not only the position of “rat” but also the position of “fat” will be strongly re-activated, leading to increased confusion errors. To the best of our knowledge, this cue-similarity principle has never been tested for lists of phonologically similar words or lists of semantically similar words.

The Present Study

The purpose of the present study is to test whether the general similarity principles introduced above also apply to semantic information. Previous studies manipulating semantic similarity have shown that people recall more semantically similar than dissimilar items (i.e., better item memory) (Poirier & Saint-Aubin, 1995). This beneficial effect is generally attributed to people using the semantic category shared by similar items to restrict the set of plausible response candidates during recall (Neale & Tehan, 2007; Saint-Aubin & Poirier, 1999b), or due to increased activation in the shared semantic network in long-term memory (Kowialiewski, Lemaire, et al., 2021; Kowialiewski & Majerus, 2020; Tse et al., 2011). We will return to the impact of similarity on item memory in the General Discussion. Whereas the evidence for improved item memory is robust, whether semantic similarity increases confusion errors is more ambiguous. Previous studies testing the impact of semantic similarity on confusion errors provided mixed results; some providing evidence for it (Baddeley, 1966; Saint-Aubin & Ouellette, 2005; Tse et al., 2011) and some providing evidence against it (Monnier & Bonthoux, 2011; Nairne & Kelley, 2004; Neale & Tehan, 2007; Saint-Aubin & Poirier, 1999b). A recent meta-regression study suggested that semantic similarity increases order errors (Ishiguro & Saito, 2020). This meta-regression is, however, not completely conclusive, as the – marginally significant – results pertain to a specific measure of semantic similarity. When a different measure of semantic similarity was used, no impact of semantic similarity was observed in the Ishiguro & Saito meta-regression study. These contradictions raise the question of whether semantic information is bound to contexts in the same way as phonology.

In addition to resolving this empirical uncertainty, we also provide a first test of a new prediction from the WM architecture presented in **Figure 1**: People should be able to retrieve a context when presented with an item and confusion errors should come from the item similarities. According to the cue-similarity principle, confusions errors should increase when similar items are used as cues to retrieve the positions. The cue-similarity principle has never been tested with this direction of retrieval in verbal WM tasks, despite being a core prediction from positional models of WM.

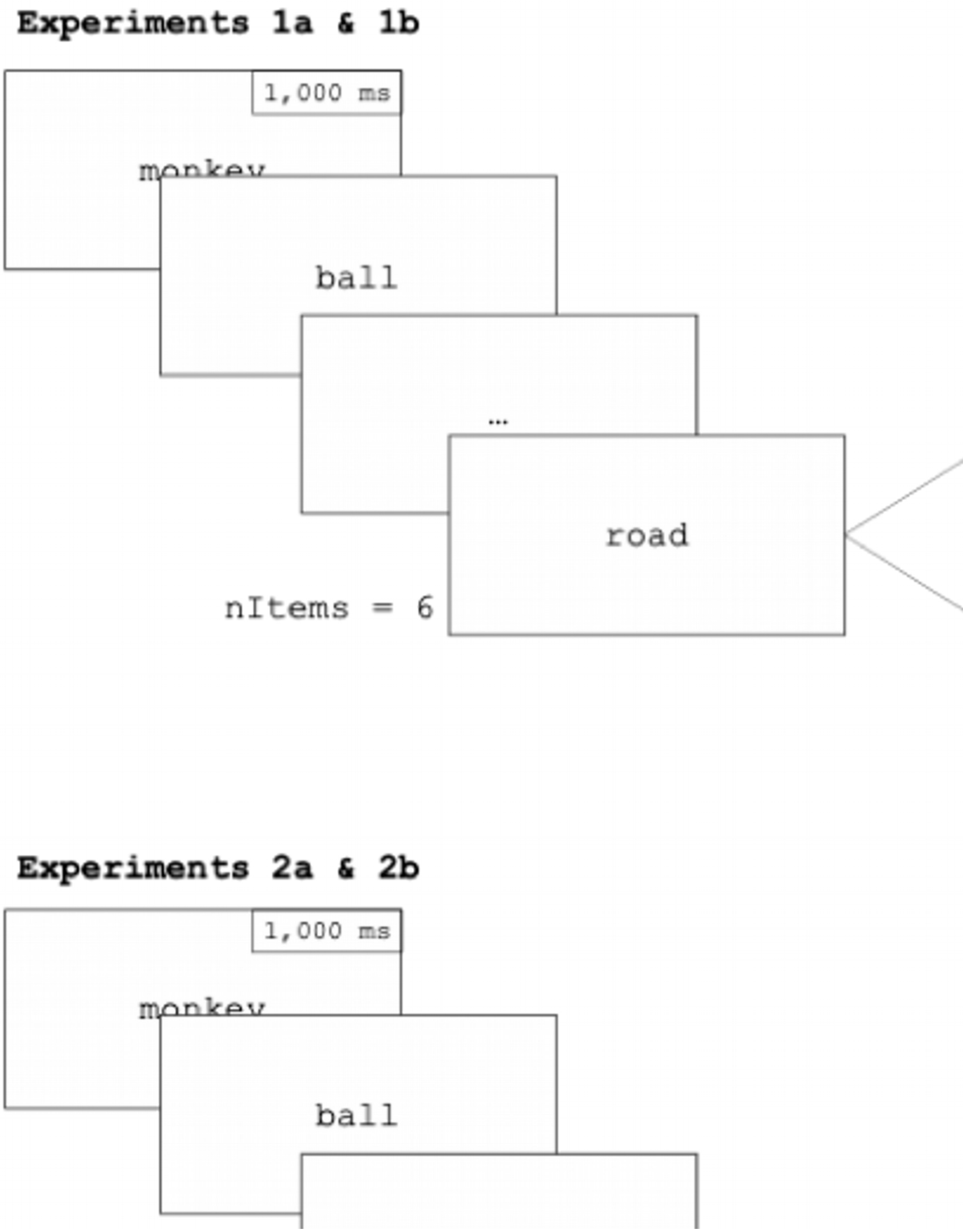
We tested whether semantic information is encoded in the same way as phonological information, namely by binding that information to appropriate context cues such as positions. Across six experiments, we manipulated semantic (Experiments 1a, 2a & 3a) and phonological (Experiments 1b, 2b & 3b) similarity between items. We used category membership to manipulate semantic similarity (e.g., musical instruments, animals, fruits), based on the assumption that similarity will be very high between members of the same category (e.g., "leopard-lion-cheetah"), compared to items drawn from different categories (e.g., "jacket-tree-letter")[1]. The phonological manipulation served as a control to assess the validity of our experimental procedures. As a rough equivalent of semantic similarity, we manipulated phonological similarity by using lists of items drawn from the same rhyming category (e.g., "rat,

[1] Category membership is a robust and safe way to study semantic similarity. The categories can be directly used to create the similar lists. The dissimilar lists are then created by sampling one word from different categories. This way, all individual characteristics of the stimuli affecting WM performance are controlled for, such as word frequency, imageability/concreteness, or neighborhood density.

fat, mat”) and compared these lists to lists of non-rhyming items. Both similarity manipulations involve categories that the similar items share (e.g., a semantic or rhyming category), and have been shown to increase the number of items people can recall (Gupta et al., 2005; Poirier & Saint-Aubin, 1995). Therefore, the two similarity manipulations are comparable. Participants were asked to bind the study list items either in relation to a temporal context (Experiment 1 & 2) or a spatial context (Experiment 3). If semantic information is bound to context the same way as phonological information, we should observe more confusion errors in semantically similar than semantically dissimilar lists.

Figure 3

Illustration of the Procedure Used Across Experiments.


Experiments 1a & 1b
1,000 ms
monkey
ball
...
road
nItems = 6
Experiments 2a & 2b
1,000 ms
monkey
ball

Note. Exp. 1a & 1b: six items appeared sequentially in the middle of the screen for 1,000 ms each. At retrieval, participants were either asked to perform a serial recall or order reconstruction task. Exp. 2a & 2b: six items appeared sequentially in the middle of the screen for 1,000 ms each. At retrieval, participants were sequentially cued with the positions and had to recall the items (cued recall of words) or were sequentially cued with the items and had to recall the positions (cued recall of positions). Exp. 3a & 3b: five items appeared sequentially on the screen on an invisible circle for 750 ms. Each word was preceded by a dot presented during 250 ms, indicating the exact center of the to-be-remembered word on the screen. On half the trials, participants were cued with a spatial location and were required to type the word associated to it (cued recall of words). On the other half, they were cued with a word and were required to report its spatial location (cued recall of spatial locations). The retrieval direction associated with each task is indicated on the right side.

The novelty of our study was to test the similarity principle across both retrieval directions. We tested the impact of the similarity manipulations on confusion errors by cueing with the context to access the items (Experiments 1, 2 and 3, see **Figure 3**), as classically done in the majority of studies. Critically, we also tested item-context binding by cueing with the items to access the context (Experiments 2 and 3, see **Figure 3**). Taking both directions of retrieval together provides an exhaustive and unambiguous test of whether semantic information is bound to contexts the same way as phonological information.

Experiments 1a & 1b

Experiments 1a & 1b assessed the impact of semantic and phonological similarity with similar vs. dissimilar lists in a serial recall and an order reconstruction task (see **Figure 3**, upper panel). In the serial recall task, participants had to retrieve the items, given positional cues, by typing the words in a prompt box. The serial recall task provides a way to assess the impact of both similarity manipulations on item and confusion errors. In the order reconstruction task, the items were given at retrieval and participants had to put them in their original order, thus providing a pure measure of order memory. If semantic information is used during item-context binding, we predict that people should confuse semantically similar items more often than dissimilar items.

Data availability

All the materials, codes, data, and data analyses across all experiments have been made available on the Open Science Framework: https://osf.io/tpsg2/

Methods

Participants. Young adults aged between 18 and 35 years participated in Experiments 1a & 1b (N = 60 for each experiment). Sample sizes were first estimated based on previous studies investigating the impact of semantic and phonological similarity, leading to a base sample size of 30. In case the Bayes Factor (see statistical procedure) did not reach a sufficient level of evidence (BF > 10 for either the null or the alternative hypothesis) concerning the critical effects of interest, thirty more participants were recruited. Sixty participants per experiment was set as the maximum N due to financial constraints.

Participants were recruited on the online platform Prolific. All participants were English native speakers, reported no history of neurological disorder or learning difficulty, and gave their written informed consent before starting the experiment. The experiment has been carried out in accordance with the ethical guidelines of the Faculty of Arts and Social Sciences at the University of Zurich.

Materials. The stimuli in the semantic experiment were drawn from 40 (Experiment 1a) and 42 (Experiment 2a & 3a) different taxonomic categories. Examples of semantic categories involved body parts (hand, foot, leg, knee, arm, elbow), vehicles (bus, car, truck, motorcycle, taxi, scooter) or emotions (happiness, love, sadness, anger, fear, disgust). In the phonological experiments (i.e., Experiments 1b, 2b & 3b), the stimuli were drawn from 42 different rhyming categories. The rhyming categories involved both monosyllabic (e.g., fain, gain, main, pain, rain, bane) and disyllabic (e.g., bangle, dangle, jangle, wangle, mangle, spangle) words. In Experiments 1a, 1b, 2a and 2b, six items per categories were included. Experiments 3a & 3b used the same categories, but only five items were used to achieve reasonable performance level, as informed by a pilot study.

We decided to draw the items from clearly defined semantic and rhyme categories to maximize between-item similarity across both the semantic and phonological dimensions. Furthermore, both similarity manipulations have in common that they increase item memory (Gupta et al., 2005; Poirier & Saint-Aubin, 1995). As the phonological manipulation served as a control to draw conclusions about the semantic manipulation, it is important to show that they have comparable effect on one dependent variable – in our case, item memory. The full list of stimuli is available on OSF. To form a similar list, six items were drawn from the same category. For each list, the six items were randomly drawn from the category, and their order was shuffled. The dissimilar lists were built by randomly sampling one item from each of the six different categories. Constraints were imposed when creating the dissimilar lists across both similarity dimensions to ensure that idiosyncratic aspects of the lists would not lead to spurious effects (see **Appendix A**).

Several metrics of semantic similarity have been proposed in the literature. Among these, Latent Semantic Analysis (LSA) is the most used. It measures the extent to which two words co-occur within similar contexts in large corpora (Landauer & Dumais, 1997). A recent study found that another variable, WordNet path length, predicts WM performance more accurately than LSA (Ensor et al., 2021). This variable measures the shortest path length that separate concepts in a hypothetical semantic network. Finally, another semantic similarity metric has been proposed, which contrary to classical measures, is thought to be partially independent from lexical connectivity measures such as LSA or WordNet path length (Ishiguro & Saito, 2020). This metric relies on three main dimensions: valence, arousal, and dominance (Moors et al., 2013). With this metric, similarity at the list-level is obtained by first computing the centroid of list items in the semantic space. The mean Euclidean distance of all items from their centroid is then computed. The closer the items from their centroid, the more similar they are. We used these metrics (i.e., LSA, WordNet path length and mean distance from the centroid) to evaluate the extent to which the similar and dissimilar lists we used differed in terms of semantic similarity. Overall, the semantically similar lists differed from the semantically dissimilar lists across all semantic similarity measures explained above. In contrast, the phonologically similar and dissimilar lists did not credibly differ along any dimension. The results from this analysis are reported in **Table 1**.

Table 1

Similarity Measures Across Similarity Manipulations

Similarity	Metric	Condition	Mean	BF_{10}
Semantic	LSA	Similar	0.264	> 100

		Dissimilar	0.003	
	WordNet Path Length	Similar	5.011	> 100
		Dissimilar	10.652	
	VAD	Similar	1.007	> 100
		Dissimilar	1.438	
Phonology	LSA	Similar	0.016	0.837
		Dissimilar	0.009	
	WordNet Path Length	Similar	9.26	0.128
		Dissimilar	9.467	
	VAD	Similar	1.299	0.487
		Dissimilar	1.326	

Procedure. The goal of these two first experiments was to provide a comprehensive direct comparison of semantic and phonological similarity effects on confusion errors, here measured through the ability to report the items in their serial order (i.e., order memory), as classically done in the serial recall literature. The items were words, and their context was the word's ordinal position in the list. The task is illustrated in **Figure 3**, upper panel. Each trial began with a central fixation point presented for 500 ms, followed by the presentation of the study list. Study lists consisted of six words presented sequentially at the center of the screen in Courier font. Each word was presented on screen for 1000 ms, followed by the next word with no inter-stimulus interval. Directly after the presentation of the last item, the retrieval phase began. On half the trials, participants were asked to perform serial recall. When this occurred, a prompt box appeared at the center of the screen, and participants were asked to type each word in the order in which they appeared. To validate each response, they pressed “Enter”. To help participants keep track of the within-list position, each prompt box was associated with a number at the bottom of it, starting from “1”. If participants did not know a given item, they were invited to leave the prompt box empty and move on to the next item, resulting in an omission error. On the other half of the trials, participants were asked to perform order reconstruction. When this occurred, the six words appeared again on the screen on a single line in a pseudorandom order. Using their computer mouse, participants sequentially clicked on each item to reconstruct the order in which the words had appeared at encoding. After each click, the selected word was replaced by a string of “#” characters. This was done to ensure that each word was discarded from the competition after being selected. Participants performed four training trials (i.e., two in each recall condition) before beginning the main experiment.

The purpose of this experimental procedure was to test the impact of similarity on memory for order in a more controlled way to what has previously been done. Both the serial recall and order reconstruction procedures require the disambiguation of WM traces by comparing them to a set of candidates. In serial recall, these candidates are the items stored in long-term memory. In order reconstruction, the candidates are the list items provided at retrieval. The type of recall test (i.e., serial recall, order reconstruction) was not revealed before the retrieval phase and was pseudo randomly assigned to each trial. This procedure ensured that the lists were encoded in the same way for each recall type, an aspect which has rarely been controlled in previous similarity manipulations. The order-reconstruction task has the advantage of providing a pure test of order memory, as item errors are impossible. The serial recall task prevented participants from memorizing only the first letter of each word, a strategy that would be successful for the order reconstruction task and would have neutralized the similarity manipulations. Instead, each item needed to be encoded as a whole to achieve reasonable recall performance in the serial recall task.

In sum, there were four different experimental conditions: two recall procedures (serial recall, order reconstruction) and two similarity conditions (similar, dissimilar). There were 20 and 21 trials for each experimental condition in Experiments 1a and 1b, respectively.

Scoring procedure. Different scoring procedure reflecting different aspects of WM were computed. First, we computed participant's ability to recall the identity of the items in the memory list. Second, we computed participant's ability to recall the items in their correct position. As only the latter is theoretically relevant for item-context binding, it was particularly important to measure it in a way that is not confounded by item memory.[2] In the following paragraphs, we explain in more details how they were computed.

In the serial recall task, we first computed an *item recall* score, for which an item was considered correct if recalled, regardless of the position at which it was output at retrieval. For instance, given the target sequence "Item1 – Item2 – Item3 – Item4 – Item5 – Item6" and the recalled sequence "Item1 – Item3 – blank – Item5 – blank – Item6", Item1, Item3, Item5 and Item6 would be considered as correct. This criterion, also illustrated in **Figure 1**, measures the ability to recall item identity. Second, we computed order memory, as the proportion of items recalled at their correct position out of the number of items recalled regardless of their position. This proportion, also illustrated in **Figure 1**, was computed by first coding all items not recalled at all as missing values, and then averaging for each participant the number of items correctly recalled in correct order at each serial position. These scores are equivalent to the order recall

[2] Researchers traditionally report the proportion of items recalled in correct position for serial recall tasks. This score has the disadvantage to provide a blend of both item and item-context binding and is therefore ambiguous regarding which aspect of WM is affected by a given manipulation. It was therefore not included.

score usually used to assess the impact of experimental manipulations on memory for order information (Saint-Aubin & Poirier, 1999a). One problematic aspect with this measure is that it depends on items being recalled at all; items not recalled can’t provide any information regarding order memory. The order reconstruction task solves this potential issue.

In the order reconstruction task, participants are asked to reconstruct the order of the to-be-remembered items. Accuracy is measured as the proportion of items chosen in their correct ordinal position. As maintenance of item information is not required in this task, reconstruction accuracy provides an unambiguous measure of the extent to which participants remember their order.

Data analysis. We conducted Bayesian analyses using the BayesFactor package (Morey & Rouder, 2014) implemented in R. Evidence in favor of a model over a comparison model is given by the Bayes Factor (BF). It reflects the likelihood ratio of a given model relative to a competing model, for instance the null model. The BF_{10} is used to denote the likelihood ratio for the alternative model relative to the null model, and the BF_{01} to denote the likelihood ratio for the null model relative to the alternative model. We use the classification of strength of evidence proposed in previous studies (Jeffreys, 1998): a BF of 1 provides no evidence, $1 < BF < 3$ provides anecdotal evidence, $3 < BF < 10$ provides moderate evidence, $10 < BF < 30$ provides strong evidence, $30 < BF < 100$ provides very strong evidence, and $100 < BF$ provides extreme/decisive evidence. In the main analyses of Experiments 1 through 3, each effect of interest was tested using a Bayesian paired-samples t-test using the aggregated data (i.e., data averaged for each participant) as dependent variable. We also report the 95% Bayesian Credible Intervals using the highest density intervals of the sampled posterior distribution of the model under investigation (number of iterations = 10^5). We used the default medium Cauchy prior distribution with scale ¿ $\frac{\sqrt{2}}{2}$. On each graph, we report the 95% within-subject Confidence Intervals for each mean.

Results

Detailed statistical values across all experiments are reported in **Table 2.**

Serial recall. As can be seen in **Figure 4** left panels, similar items were recalled more often than dissimilar items as shown by better item memory accuracy, and this difference was supported by decisive evidence for both the semantic (BF_{10} = 5.47e+19) and the phonological dimensions (BF_{10} = 1.464e+13). In contrast to item memory, confusion errors did not behave the same way across the semantic and phonological dimensions (see **Figure 4**, middle panels). As expected, phonologically dissimilar items were recalled more often in their correct order than phonologically similar items, and this difference was supported by decisive evidence (BF_{10} = 8.675e+14). Hence, phonological similarity increased confusion errors. However, semantic similarity did not influence participants' ability to recall the words in their correct order, and hence had no influence on confusion errors. This absence of an effect was supported by moderate evidence (BF_{01} = 7.035).

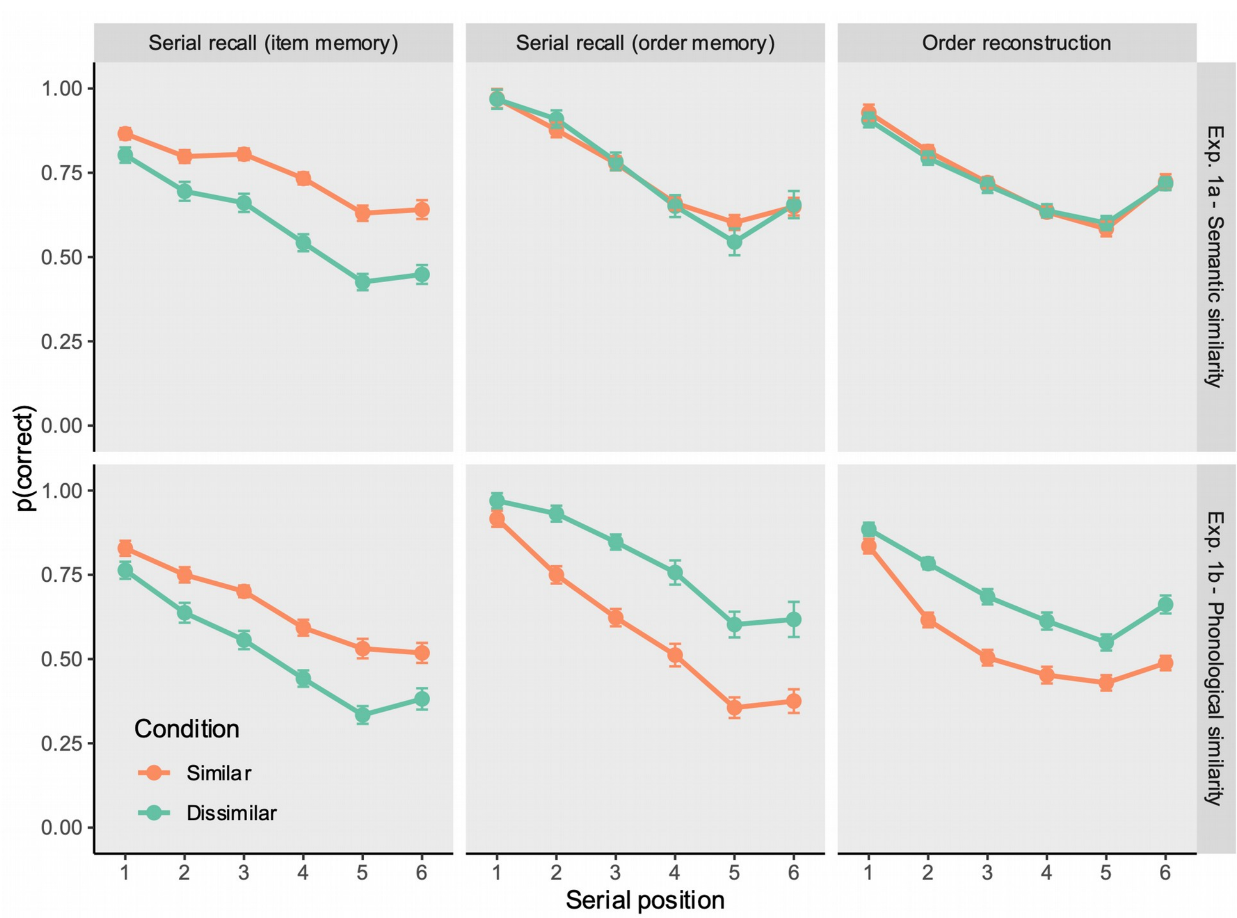
Order reconstruction. In the order reconstruction task, there was no obvious increase of confusion errors for semantically similar over dissimilar lists (see **Figure 4**, upper right panel), and moderate evidence supported this absence of a difference (BF_{01} = 6.321). In contrast, confusion errors increased for phonologically similar vs. dissimilar lists of items (see **Figure 4**, bottom right panel), and this difference was supported by decisive evidence (BF_{10} = 4.07e+8).

Discussion

Whereas both semantic and phonological similarity increased the number of items people were able to recall to about the same degree (see **Figure 4**, left panels), only phonological similarity credibly and consistently impacted confusion errors (see **Figure 4**, middle and right panels). These results replicate previous results showing a null impact of semantic similarity on memory for order (e.g., Saint-Aubin & Poirier, 1999b). In the next experiments, we tested the impact of similarity on item-context binding in a more exhaustive manner, by testing both retrieval directions.

Figure 4

Results of Experiment 1a – Semantic Similarity Manipulation (upper panel), and 1b – Phonological Similarity Manipulation (lower panel).



Note. Left panel: Serial recall (item memory). Middle panel: Word recall (order memory). Right panel: Order reconstruction. Error bars represent 95% confidence intervals for within-subject comparisons.

Experiments 2a & 2b

Experiments 2a & 2b assessed binding memory between items and ordinal-position contexts, as Experiments 1a & 1b (see **Figure 3**, middle panel). Here we also varied the direction of retrieval: Participants were presented with a position and had to retrieve the items associated to it (i.e., word recall task, context-to-item retrieval direction), or presented with an item and had to retrieve the position associated to it (i.e., position recall task, item-to-context retrieval direction). As for Experiments 1a & 1b, we predicted that semantic similarity increases confusion errors if semantic information was bound to context the same way as phonological information.

Methods

Participants. Young adults aged between 18 and 35 years participated in Experiments 2a & 2b (N = 60 for each experiment). Participants were recruited on the online platform Prolific. All participants were English native speakers, reported no history of neurological disorder or learning difficulty, and gave their written informed consent before starting the experiment. The experiment has been carried out in accordance with the ethical guidelines of the Faculty of Arts and Social Sciences at the University of Zurich.

Material. All materials were identical to those used in Experiments 1a & 1b.

Procedure. Experiments 2a and 2b used the same design as Experiments 1a & 1b, but with two new test procedures: Cued recall of words, given positions, and cued recall of positions, given words. Whereas the cued recall of words requires the retrieval of the words from the positions, the cued recall of positions requires the retrieval of the positions from the items. Each position/item were probed in a random order at retrieval. For instance, the to-be-remembered sequence “freeze, love, puma, artwork, tree, venus” could be probed such that “artwork” had to be retrieved first, followed by “venus”, then “freeze”, etc. Trials with item cues and trials with position cues were intermixed randomly so that the kind of test was not predictable during list encoding. The task is illustrated in **Figure 3**, middle panel. In the cued recall of positions task, participants were presented with a word below a prompt box and were asked to report the serial position at which the word was presented. The recall procedure continued until all positions were probed. The cued recall of words was identical to the cued recall of positions task, except that a number served as cue to retrieve the associated word. The number was presented below the prompt box, indicating the position of the to-be-recalled item.

The novel aspect of Experiments 2a & 2b is the cued recall of positions task, which induces the retrieval direction from item to context. As each position was probed in a random order independent of the order of presentation, this task discouraged participants from mentally recalling the list serially to retrieve the position. This contrasts with the typical serial recall and order reconstruction tasks, in which the retrieval direction from context to item is the most plausible strategy to perform the task. Experiments 2a and 2b manipulated semantic and phonological similarity, respectively. There were again four different experimental conditions: two recall procedures (word recall, position recall) crossed with two similarity conditions (similar, dissimilar). There were 21 trials for each experimental condition in each experiment.

Scoring procedure. For recall of the words from position (i.e., cued recall of words), similar item and order scores were used as those in Experiments 1a & 1b. For recall of the positions from words (i.e., position recall task), performance was analyzed by computing the proportion of positions correctly reported for each cued word. Note that in this task, participants produced a small number of omissions. When this occurred, the observation was treated as missing data to match more closely the order reconstruction and spatial location tasks (cf. Experiments 3a & 3b) in which omission errors are not allowed.

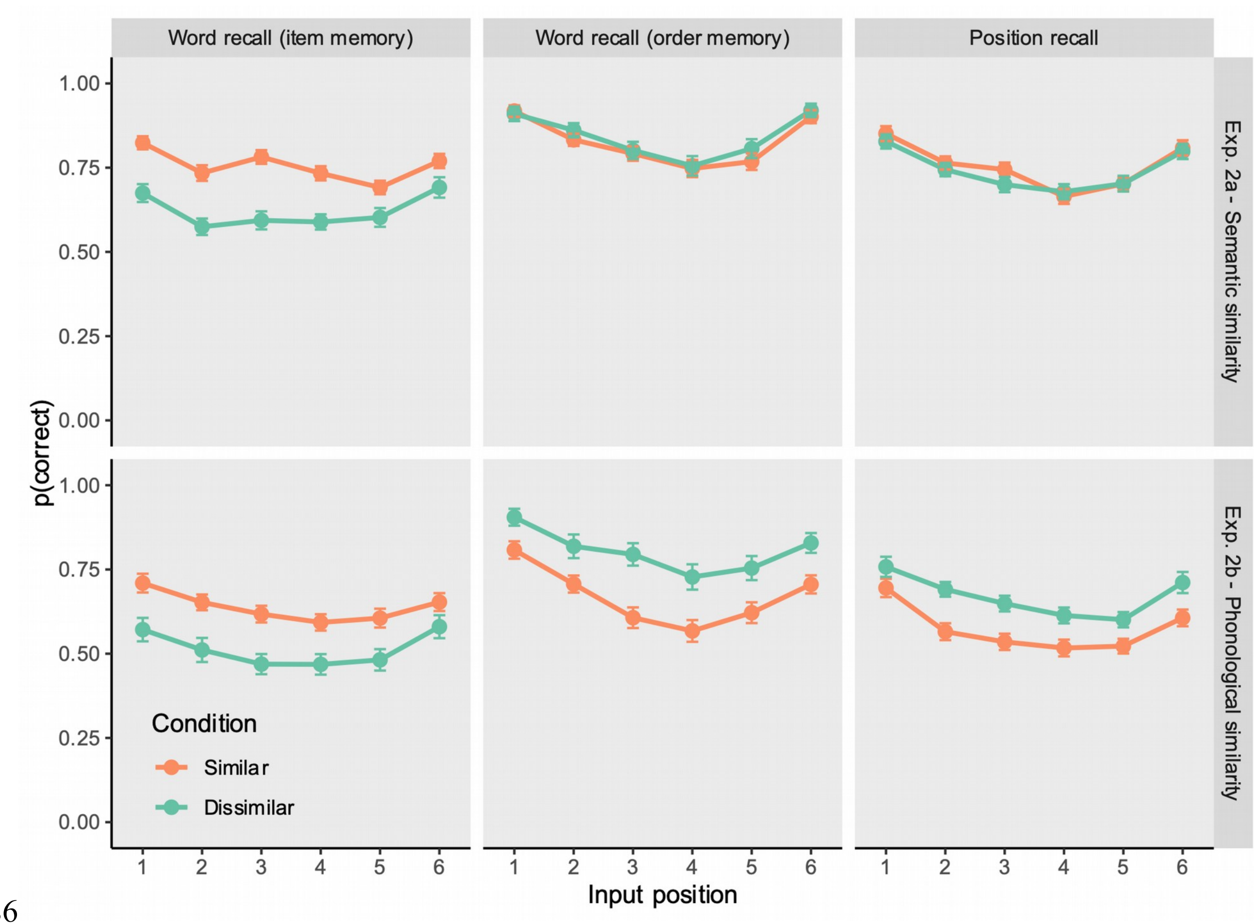
Results

Word recall. As can be seen in **Figure 5**, left panels, the results replicate those of Experiments 1a & 1b. Participants recalled more items in the similar than the dissimilar condition, and this difference was associated with decisive evidence both in the semantic (BF_{10} = 3.953e+14) and phonological (BF_{10} = 1.567e+8) dimensions. Along the phonological dimension, participants recalled the dissimilar items more often than similar items in their correct order (see **Figure 5**, lower middle panel), with decisive evidence supporting this difference (BF_{10} = 3.79e+7). Hence, people confused more often the similar versus dissimilar items. In contrast, there was no credible difference (BF_{01} = 2.177) in confusion errors between semantically similar and dissimilar lists (see **Figure 5**, top middle panel).

Cued recall of positions. Performance in the cued recall of positions task was different for semantic and phonological similarity. As can be seen in **Figure 5**, upper right panel, semantic similarity did not credibly (BF_{01} = 2.924) impair participants' ability to recall the positions associated with each item. This contrasts with phonological similarity, for which participants confused the positions more often when presented with phonologically similar versus dissimilar items (BF_{01} = 1.757e+5), as can be seen in **Figure 5**, bottom right panel.

Figure 5

Results of Experiment 2a – Semantic Similarity Manipulation (upper panel), and 2b – Phonological Similarity Manipulation (lower panel).



Note. Left panel: Word recall (item memory). Middle panel: Word recall (order memory). Right panel: Cued recall of positions. Error bars represent 95% confidence intervals for within-subject comparisons.

Discussion

Semantic and phonological similarity again enhanced the number of items participants were able to recall (see **Figure 5**, left panels). However, only phonological similarity credibly increased confusion errors (see **Figure 5**, middle and right panels). The novel result of this experiment is that when the items served as cues to recall positions, phonological similarity impaired recall but semantic similarity did not. According to the cue-similarity principle, similar cues should lead to increased confusion errors compared to dissimilar cues. The absence of an effect of semantic similarity when words were used as retrieval cues forces us to conclude that the meaning of the words played no role in their use as retrieval cues.

In the following experiments, we extended these tests by changing the nature of the context to which items were to be bound, from ordinal position to spatial location. If the findings of Experiments 2a and 2b reflect how meaning is encoded into WM in general, then we should observe them for any item-context binding and not just for item-temporal context bindings.

Experiments 3a & 3b

Experiments 1 and 2 manipulated similarity between items in tasks involving the binding between items and ordinal positions as contexts. Experiments 3a & 3b tested similarity in tasks involving the binding between items and spatial locations as context. Participants were presented with items at different spatial locations, arranged on a circle (see **Figure 3**) and had to memorize each item and its location. At retrieval, they were presented either with a location or an item. When presented with a location, they had to recall the word associated to it (i.e., word recall task, context-to-item retrieval direction). When presented with a word, they were asked to report the location associated to that word on a continuous scale (i.e., spatial location task, item-to-context retrieval direction). The spatial location task enforced the retrieval direction from item to context even more strongly than Experiments 2a & 2b. As the temporal dimension was irrelevant in Experiment 3a & 3b, this further discouraged participants to rehearse the word list in its presentation order before each response. We expected to find more confusion errors in the semantically similar vs. dissimilar lists if semantic information is bound to context.

Methods

Participants. Young adults aged between 18 and 35 years participated in Experiments 3a & 3b (N = 60 for each experiment). Participants were recruited on the online platform Prolific. All participants were English native speakers, reported no history of neurological disorder or learning difficulty, and gave their written informed consent before starting the experiment. The experiment has been carried out in accordance with the ethical guidelines of the Faculty of Arts and Social Sciences at the University of Zurich.

Material. This experiment used the same words as in Experiments 2a & 2b. The number of words to be remembered was reduced from 6 to 5, as the task was slightly more difficult than the previous ones, as informed by a pilot study.

Procedure. Experiments 3a and 3b differed from Experiments 2a & 2b only in the kind of context to which the items were to be bound. The items were words, and the context was the spatial location of each word on the screen. The task is illustrated in **Figure 3**, lower panel. Participants encoded 5-item study lists, with each item being sequentially presented in lower case at a pace of 1 item/s (250 ms OFF, 750 ms ON). Each word appeared at a different location on an invisible circle centered around the middle of the screen. The locations were pseudo randomly sampled with the constraint that the angular distance (in degree) between any two locations should not be smaller than a pre-defined value (see **Appendix B** for the methodological details). To ensure that participants could correctly identify the center of each item in an unambiguous manner, the words were preceded by a dot presented during 250 ms, indicating the exact center of each item. Directly after the encoding phase, there was an interval of 1000 ms, followed by the retrieval phase. During the retrieval phase, the circle around which the items were initially presented was always displayed on the screen.

As in Experiments 2a & 2b, the items were not tested in their order of presentation, an aspect of the procedure which made the temporal dimension irrelevant. On half the trials, the participants were cued with a previously presented location on the wheel and had to recall the words associated to it by typing it in a prompt box. Participants were asked to leave the box empty if they were not able to retrieve a word. After pressing the “Enter” key, another location was cued, and this process repeated until all memoranda were tested. On the other half of the trials, a word from the to-be-remembered list appeared at the center of the screen written in uppercase. Participants were asked to report on the wheel the spatial location to which the item was associated. To help participants locate their response as accurately as possible, a dot was continuously presented on the wheel, based on the direction in which the current mouse position deviated from the screen center. To confirm their response, participants clicked on the desired location. The response automatically initiated the next retrieval attempt, until all words were tested. Participants performed four training trials (i.e., two in each recall condition) before beginning the main experiment.

There were again four different experimental conditions: two recall procedures (cued recall, spatial location reproduction) across two similarity conditions (similar, dissimilar). Twenty-one trials were included in each experimental condition in both experiments.

Scoring procedure. When participants had to recall the items from their spatial location, the same scoring procedure was used as Experiments 1a, 1b, 2a and 2b for item memory. Order memory was computed as the proportion of words recalled at their correct spatial location out of the number of words recalled regardless of their location. For the spatial location task, which involved participants reporting the word locations on a continuous circular scale, we measured the absolute angular distance (in degrees) of participant's response to the target location. We calculated the average absolute angular distance for each condition and each participant.

Results

Word recall. As can be seen in **Figure 6**, both semantic (upper left panel) and phonological (bottom left panel) similarity credibly (BF_{10} = 1.881e+18 and BF_{10} = 1.926e+9, respectively) increased the number of items recalled, with decisive evidence supporting a recall advantage for similar vs. dissimilar lists. Semantically similar items were not confused more often between each other than dissimilar lists (see **Figure 6**, upper middle panel), and an absence of difference was supported by anecdotal evidence (BF_{01} = 2.948). This result contrasts with what is observed in the phonological dimension, with phonologically similar lists being more often confused as compared to dissimilar lists (see **Figure 6**, lower middle panel). This difference was supported by decisive evidence (BF_{10} = 7.145e+4).

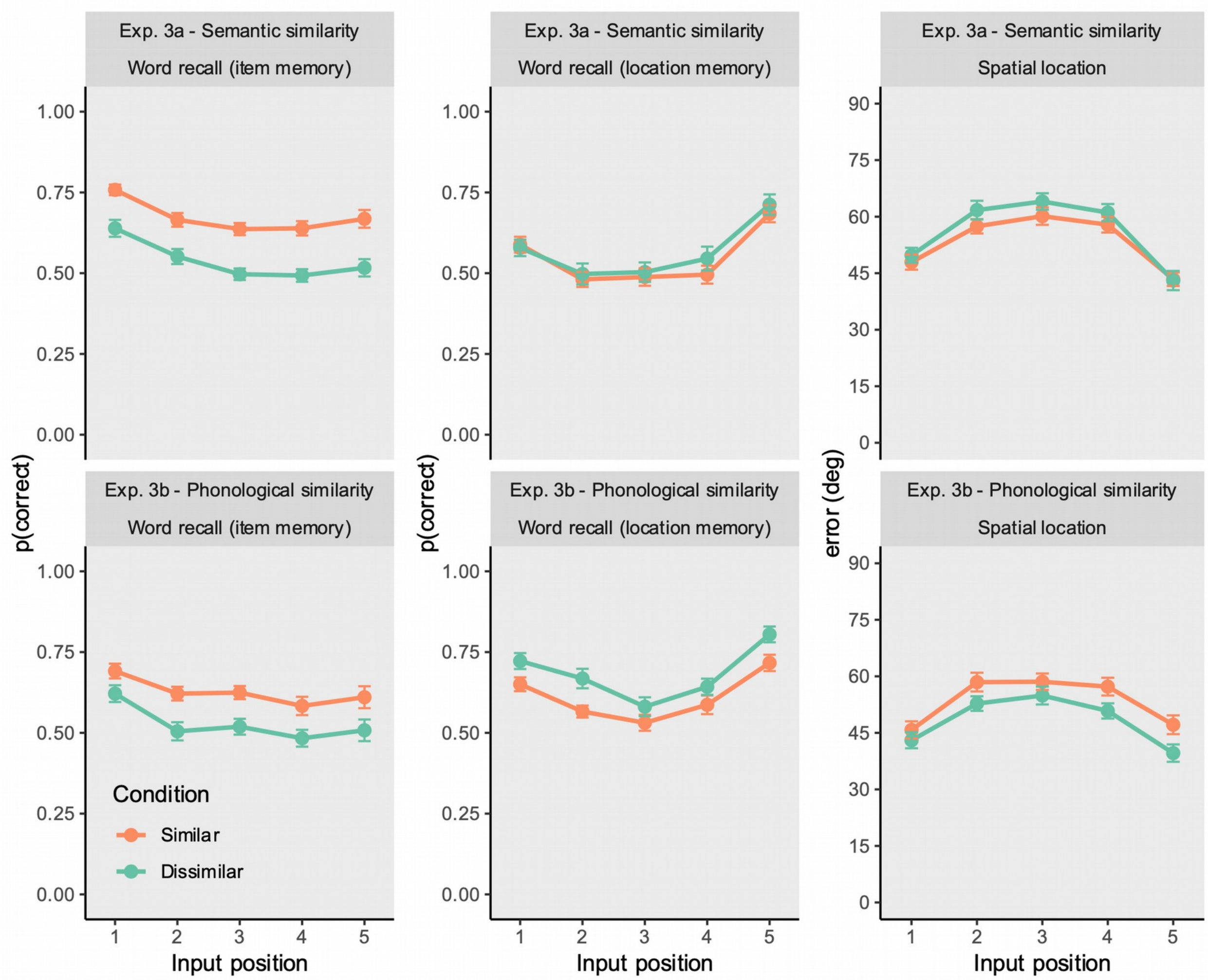
Cued recall of spatial locations. Results on the spatial location task in **Figure 6**, upper right panel, suggest that the semantically similar and dissimilar lists did not substantially differ in angular error, and only anecdotal evidence supported a difference between both semantic conditions (BF_{10} = 1.902). If anything, the direction of this difference went in the opposite direction of what would be expected if similarity led to more confusion errors. In contrast, it can be seen in **Figure 6**, bottom right panel, that phonologically similar lists were associated with higher angular error in reproducing the word's location than phonologically dissimilar lists, and this difference was associated with decisive evidence (BF_{10} = 104.162).

Discussion

The present results converge with those from the previous experiments. Whereas both semantic and phonological similarity credibly increased the number of items participants recalled, only phonological similarity increased confusion errors. The phonological similarity effect was still observed even when location was used as contexts instead of positions. To the best of our knowledge, this result has never been reported in the verbal WM literature and constitutes an important test of the generality of models in which the core process of encoding into WM is the formation of item-context bindings. In the next section, we re-analyzed our data with a continuous metric of semantic similarity recently proposed in the literature.

Figure 6

Results of Experiment 3a – Semantic Similarity Manipulation (upper panel), and 3b – Phonological Similarity Manipulation (lower panel).



Note. *Left panel: Word recall (item memory). Middle panel: Word recall (order memory). Right panel: Cued recall of spatial locations. Error bars represent 95% confidence intervals for within-subject comparisons.*

Table 2
Detailed Statistics Across all Experiments

Experiment	Task/Criterion	BF_{10}	Cohen's d	$CI_{95\%}$	Effect direction
Exp. 1a Semantic	Serial recall (item memory)	5.47e+19	2.037	[1.558; 2.449]	Sim > Dis
	Serial recall (order memory)	1/7.035	-0.015	[-0.262; -0.23]	Sim = Dis

	Order reconstruction	1/6.321	0.063	[-0.185; 0.307]	Sim = Dis
Exp. 1b Phonology	Serial recall (item memory)	1.464e+13	1.434	[1.036; 1.764]	Sim > Dis
	Serial recall (order memory)	8.675e+14	-1.587	[-1.94; -1.175]	Sim < Dis
	Order reconstruction	4.07e+8	-1.062	[-1.358; -0.722]	Sim < Dis
Exp. 2a Semantic	Word recall (item memory)	3.953e+14	1.557	[1.143; 1.899]	Sim > Dis
	Word recall (order memory)	1/2.177	-0.205	[-0.445; 0.05]	Sim ~ Dis
	Cued recall of positions	1/2.924	0.177	[-0.082; 0.416]	Sim ~ Dis
Exp. 2b Phonology	Word recall (item memory)	1.567e+8	1.029	[0.685; 1.314]	Sim > Dis
	Word recall (order memory)	3.79e+7	-0.979	[-1.258; -0.641]	Sim < Dis
	Cued recall of positions	1.757e+5	-0.792	[-1.058; -0.478]	Sim < Dis
Exp. 3a Semantic	Word recall (item memory)	1.881e+18	1.894	[1.439; 2.29]	Sim > Dis
	Word recall (location memory)	1/2.948	-0.176	[-0.413; 0.084]	Sim ~ Dis
	Cued recall of spatial locations	1.902	-0.308	[-0.548; -0.041]	Sim ~ Dis
Exp. 3b Phonology	Word recall (item memory)	1.926e+9	1.115	[0.76; 1.41]	Sim > Dis
	Word recall (location memory)	7.145e+4	-0.76	[-1.027; -0.452]	Sim < Dis
	Cued recall of spatial locations	104.162	0.508	[0.212; 0.743]	Sim > Dis

Note. For recall of locations, the dependent variable is angular error, so larger values of similar than dissimilar lists reflect poorer location memory of similar lists. The 95% credible intervals were computed at the effect-size scale.

Relationship between WM performance and the dimensional view of semantic similarity

A recent meta-regression study suggested that the absence of detrimental effect of semantic similarity on order memory might be due to an inappropriate measure of semantic similarity (Ishiguro & Saito, 2020). The authors argued that semantic similarity by category membership is confounded with relationships between concepts in a semantic network. Instead, the “true” semantic similarity between items would be better characterized by their shared features. They proposed a three-dimensional feature space encompassing valence, arousal, and dominance (Moors et al., 2013) to measure the similarity between words. The average semantic dissimilarity for a list is computed by taking the Euclidean distance for all list items from their centroid in this space. We explored whether this metric was a credible predictor of confusion errors across all our experiments manipulating semantic similarity (i.e., Experiments 1a, 2a, and 3a). We ran a Bayesian generalized mixed model with serial position and the mean distance from centroid as predictors for the recall success of each list item. Details of this new analysis are reported in **Appendix C**. We report in **Figure 7**, upper panel, the posterior distribution for all models. The results are clear-cut. There was no credible effect of the mean distance from the centroid on confusion errors. No consistent trend was observed throughout the experiments.

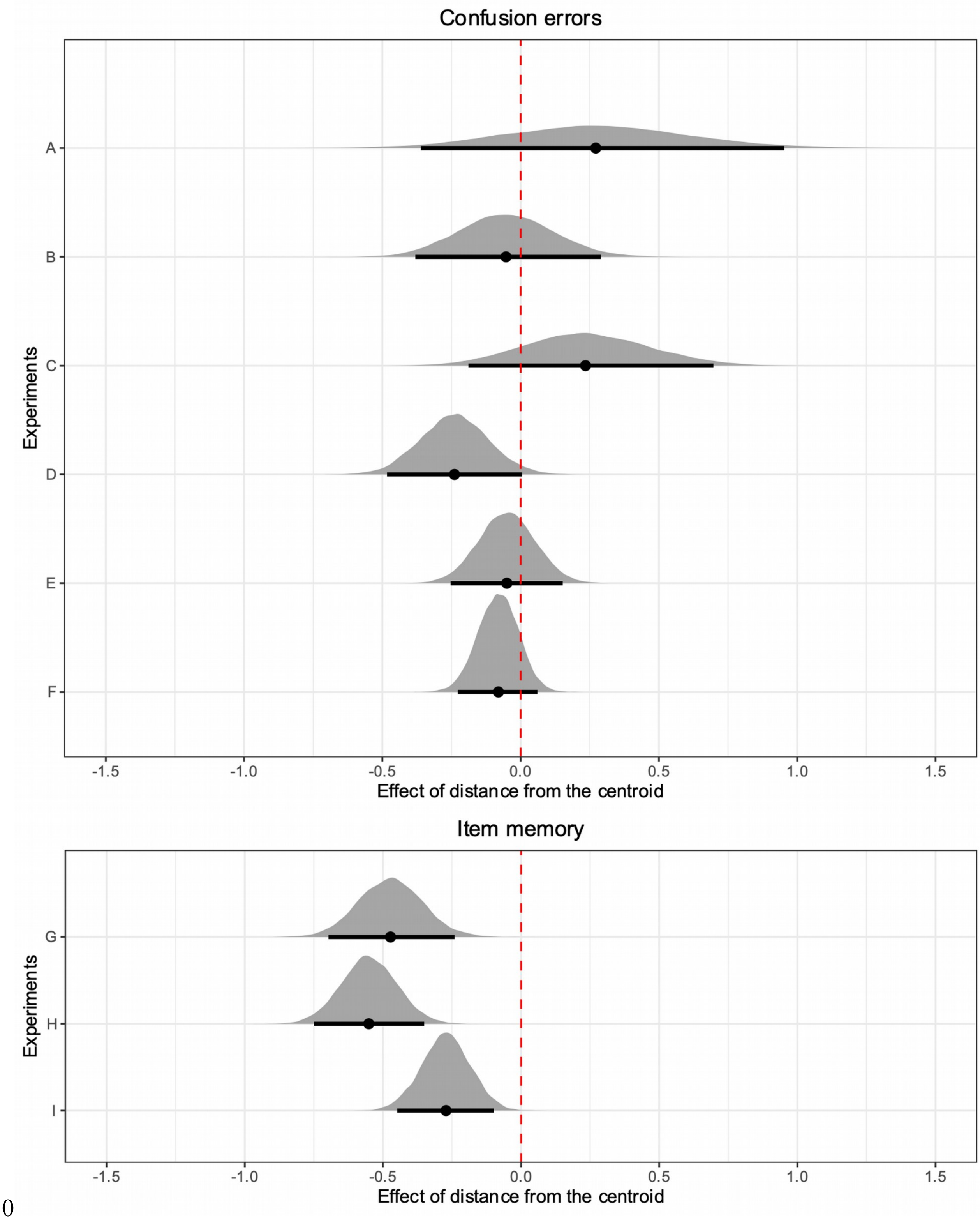
We ran similar analyses on the item-memory scores, assuming a Bernoulli distribution. The results are reported in **Figure 7**, lower panel. As can be seen, the mean distance from the centroid credibly impacted item memory consistently across Experiments 1a, 2a and 3a. As the distance from the centroid decreased (and therefore semantic similarity increased), memory for item increased. In the next section, we discuss more thoroughly the theoretical implications of these results.

Figure 7

Posterior Distributions for the Effect of Distance from the Centroid

WM SEMANTIC PHONOLOGY

629



Note. Upper panel: Confusion errors. A = Immediate recall (Exp. 1a); B = Order reconstruction (Exp. 1a); C = Word recall (Exp. 2a); D = Cued recall of positions (Exp. 2a); E = Word recall

(Exp. 3a); F = Cued recall of locations (Exp. 3a); Lower panel: Item recall. G = Immediate recall (Exp. 1a); H = Word recall (Exp. 2a); I = Word recall (Exp. 3a).

General Discussion

The present experiments yielded two main outcomes. First, both semantic and phonological similarity enhanced the ability to recall item information. Second, whereas phonological similarity credibly decreased performance in all tasks testing item-context bindings (i.e., order memory, location memory, order reconstruction, cued recall of positions, and cued recall of spatial locations), semantic similarity did not. These results provide strong converging evidence for a dissociation between phonological and semantic similarity effects in WM. Given these results, together with other empirical evidence showing an absence of semantic similarity effect on confusion errors (Neale & Tehan, 2007; Poirier & Saint-Aubin, 1995; Saint-Aubin & Poirier, 1999b), we conclude that semantic similarity does not negatively affect order and positional memory in tests of WM. If semantic information was bound to a positional or spatial context the same way as phonology, semantic similarity should have led to confusion errors, as observed for phonological similarity (Baddeley, 1966), and other dimensions of similarity (Jalbert et al., 2008; Saito et al., 2008; Visscher et al., 2007).

In the present work, we focused on the item-context binding process of WM. Based on this definition of encoding features into WM, we conclude that WM does not bind meaning to context in the same way as phonology. Other theoretical and modelling approaches would logically reach the same conclusion. For instance, in the Feature Model (Nairne, 1990) as well as its revised version (Poirier et al., 2019; Saint-Aubin et al., 2021), items are represented by vectors of perceptual and/or internally generated features. At retrieval, items stored in primary (short-term or working) memory need to be compared to items in secondary (long-term) memory. Similarity in this model leads to increased confusions because the traces in primary memory will be less discriminable when comparing them to items stored in secondary memory. Likewise, in the temporal distinctiveness account (Brown et al., 2007), similarity is computed as the Euclidean distance between items represented in a multidimensional space (e.g., temporal, phonological). The closer the items in this Euclidean space, the more confusable they are. For all these models, adding the assumption that semantics is represented in WM in the same way as phonology would necessarily result in increased confusion errors for semantically similar vs. dissimilar items, in contrast with our results.

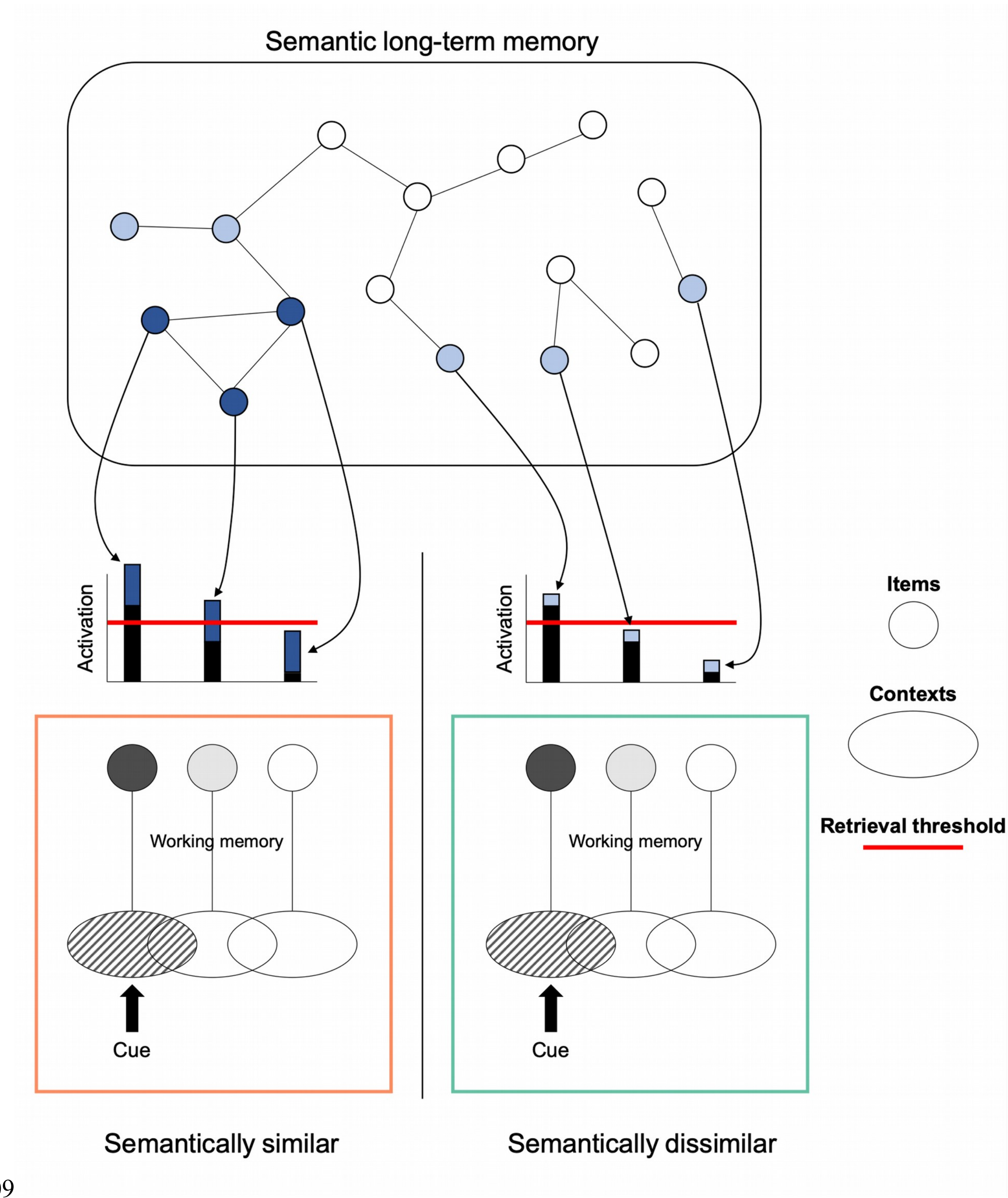
Implications for Models of Working Memory

Based on our results, we propose that semantics does not contribute to WM through the binding of semantic features to context the same way as phonology. One possibility to explain these results is to assume that semantic information is not bound to context at all. How can we explain the recall advantage for semantically similar vs. dissimilar words at the item level, if semantic information is not bound to contexts? There is robust evidence showing that semantic knowledge strongly contributes to WM performance (see Kowialiewski & Majerus, 2020 for a short meta-analysis in serial recall), with lists of semantically similar items being better recalled than lists of dissimilar ones. Results from the present study converge with these observations. The recall advantage for semantically similar vs. dissimilar items can be explained by assuming that WM partly relies on activated long-term memory, as assumed in an embedded processes account of WM (Cowan, 1999; Dell et al., 1997; Majerus, 2013; Nee & Jonides, 2013; Oberauer, 2002, 2009). Accordingly, the encoding of an item activates its long-term memory representation, including its meaning. We illustrate in **Figure 8** the mechanistic principles behind this idea. Semantically related items reactivate each other, either via their shared semantic features (Dell et al., 1997) or via lateral excitatory connections (Hofmann & Jacobs, 2014). For instance, when encoding the word “piano”, the word “guitar” would in turn be activated (Collins & Loftus, 1975). Thereby, semantically similar list items have increased activation in the semantic network. In many computational models of WM, the success in recalling an item at all depends on its ability to overcome a retrieval threshold. If an item’s activation is below the threshold, the model produces an omission. Accordingly, the higher activation of semantically similar items would help them to overcome this retrieval threshold more often than dissimilar items, leading to a recall advantage for semantically similar vs. dissimilar items which is restricted to item memory. The model presented in **Figure 8** furthermore assumes that semantic features are not directly bound to contexts. This simplifying assumption leads to an absence of a semantic similarity effect on confusion errors.

Such a model, inspired by embedded processes models of WM, helps to explain the presence of false memories in WM tasks (Abadie & Camos, 2019; Atkins & Reuter-Lorenz, 2008). When presented with a list of semantically similar items such as “leopard, tiger, lion, cheetah”, people are more likely to respond “old” in a recognition test when presented with a semantically similar lure, such as “puma”, than for a dissimilar lure, such as “desktop”. This result can be explained by assuming that activation spreads to similar list items and non-list items. Hence, non-list items that are similar to several or all list items will be activated to some extent (see **Figure 8**, similar condition). When presented with a semantically similar lure (i.e., “puma”), people are therefore more likely to say that this item was presented in the list (i.e., responding “old”), because it is now more strongly activated than other dissimilar lures (i.e., “desktop”). In contrast, when people are presented with lists of semantically dissimilar items (e.g., “arm, tree, sofa, mouse”), no such false memories are observed (Cowan et al., 2022). From the model presented in **Figure 8**, this latter result is predicted, because when given a dissimilar list, the activation spreading from list words no longer converges on the same non-list words (see **Figure 8**, dissimilar condition).

Figure 8

A Model of the Semantic Similarity Effect



Note. *When encoding items into WM, a new binding is created between this item and its context. At the same time, this item becomes activated in semantic long-term memory. Semantically similar items are assumed to have direct connections in the semantic network and spread* *activation to each other. When trying to retrieve an item by cueing it with its context, this item has an activation level, which is a combination of the activation provided by the item's binding to its context and its activation in semantic memory. If the activation level of the item is beyond a retrieval threshold, it is recalled. Otherwise, an omission is produced. When semantically similar items are encoded in the same list, they have a higher activation level thanks to the spreading of activation principle, which helps them to overcome the retrieval threshold more often than semantically dissimilar items.*

Alternative Explanations

An alternative explanation of the lack of semantic similarity effect on confusion errors is that semantic information is bound to contexts, but for some reasons, is immune to confusion errors. The only piece of evidence supporting the idea that semantic is bound to contexts comes from Kowialiewski et al. (2021). They observed than semantic knowledge can *constrain* the processing of serial order information. They presented lists composed of two semantically similar triplets (e.g., "leopard, lion, cheetah, arm, elbow, leg"). When items are recalled in a wrong position, they tend to stay within their group of similar items, rather than move to positions that have been occupied by dissimilar items, compared to the same positions in a completely dissimilar list. These results are difficult to explain without assuming that at least some form of meaning is bound to contexts. Meaning could use a different representational format, such as sparse distributed representations (Kanerva, 1988). Using a sparse code for items' meaning would prevent semantically similar items from being confused with each other, while still allowing the cognitive system to have some information about which semantic category was in which list position.

However, Kowialiewski et al. (2021)'s results can also be explained by assuming that people augment the positions of semantically similar items with a shared positional context. A similar assumption is already made in positional models to explain temporal grouping effects in serial recall (Burgess & Hitch, 1999; Henson, 1998). If semantic groups are represented like temporal groups, semantically similar items would be associated with similar positional contexts. This leads to the prediction that transposition errors should occur more often between items from the same (semantic or temporal) group than with items from another group. This explanation doesn't require semantic information to be bound to contexts.

Finally, it could be argued that the absence of a semantic similarity effect on memory for order is due to semantic knowledge not being activated in our WM task, perhaps because it needs more time to be activated. This explanation is unlikely for the following reasons. First, access to meaning is an automatic and extremely fast process, especially in language (Cheyette & Plaut, 2017; Potter, 1976; Potter et al., 2014; Tyler et al., 2002). Second, the fact that we observed very strong beneficial effects of semantic similarity on item memory goes against this claim. It shows that people had access to words' meaning and used it to increase the number of items they could recall. Strong semantic similarity effects can even be observed in running span procedures using fast presentation of memoranda (Kowialiewski & Majerus, 2018).

Possible Limitations

One possible objection to our interpretation is that phonological and semantic similarity measurements were not equivalent. This is unlikely because both kinds of similarity manipulations led to comparably strong impact on item memory, showing that people were able to detect the presence of similarities to about the same extent across both manipulations. In addition, strong phonological similarity effects on order memory can be already observed with much weaker manipulations than ours, for instance when lists items share only one phoneme (Camos et al., 2013; Fallon et al., 2005; Gupta et al., 2005). Furthermore, we are confident that our semantic similarity manipulation was a robust one, as our similar and dissimilar lists strongly differed across several semantic-similarity metrics (see **Table 1)**. If the item-context binding process was subject to confusion errors driven by semantic similarity, we would have expected at least small detrimental effects on memory for order.

It is also possible that the measures we used for item and confusion errors do not reflect what we wanted to measure. For instance, it has been argued that order reconstruction is not a pure measure of confusion errors, and could also partially reflect item memory (Neath, 1997). Contrary to this latter claim, three main outcomes support the validity of our measures. First, none of the semantic manipulations affected confusion errors, despite strongly affecting item memory. If our measures of confusion errors were not process pure, they should have been affected by semantic similarity in one way or another. This was not observed. Second, the rhyming manipulation led to a dramatic drop of performance on order memory, despite strongly enhancing item memory. If our confusion-error measures were also affected by item information, we shouldn’t have observed these divergent effects of phonological similarity on memory for item and confusion errors. Finally, all measures of confusion errors converged toward the same pattern of performance. The results illustrated in **Figure 3**, middle and right panels, clearly indicate similar performance level and serial position curves across all experiments and similarity manipulations. We can therefore be confident that all our measures of confusion errors reflect the same construct.

Conclusion

To sum up, we tested how phonological and semantic similarity impacted the maintenance of novel item-context bindings in WM. Our exhaustive tests showed that phonological similarity increases confusions errors, leading to a performance decline in all WM tasks we used. By contrast, across all experiments, semantic similarity did not increase confusion errors and did not decrease WM performance. These results imply that there is a fundamental difference between the representation of semantics and phonology in verbal WM. Either semantics is not bound to contexts, or it is bound to contexts, but in a different way than other kinds of information, such that it does not lead to confusion errors. The benefit of semantic similarity on item memory, can be explained by assuming that semantically similar items activate each other in long-term memory through their associations in a semantic network.

Acknowledgement

The research reported in this article was supported by funding from the Swiss National Science Foundation (SNSF, grant number 100014_179002).

Appendix A

Semantic manipulations

Throughout the experiments semantic similarity was manipulated by selecting words from one category (to create similar lists) and words from different categories (to create dissimilar lists). For the dissimilar lists, some restrictions were imposed regarding the combinations of categories from which words were sampled. This was done to avoid sampling from categories that are semantically related themselves, and thereby introducing similarity between words in a dissimilar list. For instance, we avoided sampling from the category “drinks” (i.e., whiskey) if a word from the category “container” (i.e., glass) was included in the list. For that purpose, the a priori semantic relationships between the categories were identified using a confusion matrix. Because of the limited number of categories available, this restriction could not be applied in a strict manner. Consequently, we allowed for those combinations to occur a neglectable number of times (i.e., once across the whole experiment).

Phonological manipulations

Phonological similarity was manipulated by generating lists of rhyming words (similar lists) and lists of non-rhyming words (dissimilar words). The rhyming categories were partially taken from Gupta et al. (2005) and Nimmo & Roodenrys (2004) studies, and involved both monosyllabic (e.g., fain, gain, main, pain, rain, bane) and disyllabic (e.g., bangle dangle jangle wangle mangle spangle) words. To ensure that non-rhyming words are sufficiently dissimilar, the Levenshtein distance between all possible pairs of words was computed. To create the dissimilar lists, we selected words that were maximally high in Levenshtein distance from each other. Furthermore, the minimal distance between every pair of items within a dissimilar list was set to a value of two. Thereby, we only allowed for one pair to have such a low distance within the same list. All the remaining pairs were constrained to have a Levenshtein distance above the value of two to ensure sufficient phonological dissimilarity between the items in a dissimilar list.

In addition, we kept semantic similarity equal between the phonologically similar and dissimilar lists. Therefore, LSA (latent semantic analysis) values were obtained for each pair of stimuli within each list using the TASA semantic space available at the following address: https://sites.google.com/site/fritzgntr/software-resources/semantic_spaces (see also Günther et al., 2015). We then compared the LSA values between the dissimilar and the similar lists. Dissimilar lists were only included in the experiment if there was no evidence for a difference in LSA similarity between them and the similar lists. As a criterion we determined a BF superior to 3 in favor of the absence of a difference (obtained in a Bayesian independent samples t-test). If the BF was below 3, new dissimilar lists were generated until this criterion was met.

Appendix B

In Experiments 3a and 3b, all to-be-remembered locations had to have a minimal distance of 40 degrees on the wheel. This limitation was introduced to ensure that stimuli were presented in distinct spatial positions. To control for a possible influence of spatial positions, we matched the spatial positions for both similarity conditions, and both recall conditions. We therefore generated lists of spatial positions that we used in both similarity conditions. To make sure that the spatial positions are the same for both recall conditions, we randomly selected half of the lists of the dissimilar condition to occur in the “cued by word”-condition. We then selected the same half of lists (aka the same spatial positions) in the similar condition to serve for the “cued by word” condition.

Appendix C

The Bayesian generalized mixed models were run using the brms package (Bürkner, 2017) implemented in R. All models included a random intercept, the random effects of both predictors, as well as their interaction. When estimating the impact of the mean deviation from the centroid on confusion errors, for Experiment 1a & 2a, the dependent variables were unaggregated accuracies of recall of each item in its position (0 vs. 1), with models assuming a Bernoulli distribution. Items not recalled at all were scored as missing data (see scoring procedure above). For Experiments 3a, the model assumed a Von Mises distribution, as the dependent variable was the raw deviation from the target. For this model, the kappa parameter (i.e., concentration around the mean, fixed to 0) was estimated. When estimating the impact of the mean deviation from the centroid on item memory, the dependent variables were unaggregated accuracies of recall of each item recalled regardless of their output position (0 vs. 1), with models assuming a Bernoulli distribution. Parameters of the models were estimated using 4 independent Markov Chains, each with 5000 samples, including 500 warmup samples. Across all analyses, the Markov Chains always converged, as indicated by R-hat < 1.05.

References

Abadie, M., & Camos, V. (2019). False memory at short and long term. *Journal of Experimental Psychology: General*, *148*(8), 1312-1334. https://doi.org/10.1037/xge0000526

Atkins, A. S., & Reuter-Lorenz, P. A. (2008). False working memories? Semantic distortion in a mere 4 seconds. *Memory & Cognition*, *36*(1), 74-81. https://doi.org/10.3758/MC.36.1.74

Baddeley, A. D. (1966). Short-term Memory for Word Sequences as a Function of Acoustic, Semantic and Formal Similarity. *Quarterly Journal of Experimental Psychology*, *18*(4), 362-365. https://doi.org/10.1080/14640746608400055

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539-576. https://doi.org/10.1037/0033-295X.114.3.539

Burgess, N., & Hitch, G. J. (1999). Memory for Serial Order : A Network Model of the Phonological Loop and Its Timing. *Psychological Review*, *106*(3), 551-581.

Burgess, N., & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language*, *55*(4), 627-652. https://doi.org/10.1016/j.jml.2006.08.005

Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Camos, V., Mora, G., & Barrouillet, P. (2013). Phonological Similarity Effect in Complex Span Task. *Quarterly Journal of Experimental Psychology*, *66*(10), 1927-1950. https://doi.org/10.1080/17470218.2013.768275

Cheyette, S. J., & Plaut, D. C. (2017). Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition*, *162*, 153-166. https://doi.org/10.1016/j.cognition.2016.10.016

Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, *82*(6), 407-428.

Cowan, N. (1999). An Embedded-Processes Model of working memory. In *Models of working memory : Mechanisms of active maintenance and executive control* (Cambridge University Press, p. 62-101).

Cowan, N., Guitard, D., Greene, N. R., & Fiset, S. (2022). Exploring the use of phonological and semantic representations in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001077

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801-838. https://doi.org/10.1037/0033-295X.104.4.801

Ensor, T. M., MacMillan, M. B., Neath, I., & Surprenant, A. M. (2021). Calculating semantic relatedness of lists of nouns using WordNet path length. *Behavior Research Methods*, *53*(6), 2430-2438. https://doi.org/10.3758/s13428-021-01570-0

Fallon, A. B., Mak, E., Tehan, G., & Daly, Charmaine. (2005). Lexicality and phonological similarity : A challenge for the retrieval-based account of serial recall? *Memory*, *13*(3-4), 349-356. https://doi.org/10.1080/09658210344000215

Farrell, S., & Lewandowsky, S. (2004). Modelling transposition latencies : Constraints for theories of serial order memory. *Journal of Memory and Language*, *51*(1), 115-135. https://doi.org/10.1016/j.jml.2004.03.007

Gorin, S., Kowialiewski, B., & Majerus, S. (2016). Domain-Generality of Timing-Based Serial Order Processes in Short-Term Memory : New Insights from Musical and Verbal Domains. *PLOS ONE*, *11*(12), e0168699. https://doi.org/10.1371/journal.pone.0168699

Guitard, D., & Cowan, N. (2020). Do we use visual codes when information is not presented visually? *Memory & Cognition*, *48*(8), 1522-1536. https://doi.org/10.3758/s13421-020-01054-0

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun—An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930-944. https://doi.org/10.3758/s13428-014-0529-0

Gupta, P., Lipinski, J., & Aktunc, E. (2005). Reexamining the phonological similarity effect in immediate serial recall : The roles of type of similarity, category cuing, and item recall. *Memory & Cognition*, *33*(6), 1001-1016. https://doi.org/10.3758/BF03193208

Henson, R. N. A. (1998). Short-term memory for serial order : The start-end model. *Cognitive Psychology*, *36*(2), 73-137.

Henson, R. N. A., Hartley, T., Burgess, N., Hitch, G., & Flude, B. (2003). Selective Interference with Verbal Short-Term Memory for Serial Order Information : A New Paradigm and Tests of a Timing-Signal Hypothesis. *The Quarterly Journal of Experimental Psychology Section A*, *56*(8), 1307-1334. https://doi.org/10.1080/02724980244000747

Hofmann, M. J., & Jacobs, A. M. (2014). Interactive activation and competition models and semantic context : From behavioral to brain data. *Neuroscience & Biobehavioral Reviews*, *46*, 85-104. https://doi.org/10.1016/j.neubiorev.2014.06.011

Ishiguro, S., & Saito, S. (2020). The detrimental effect of semantic similarity in short-term memory tasks : A meta-regression approach. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01815-7

Jalbert, A., Saint-Aubin, J., & Tremblay, S. (2008). Short Article : Visual Similarity in Short-Term Recall for Where and When. *Quarterly Journal of Experimental Psychology*, *61*(3), 353-360. https://doi.org/10.1080/17470210701634537

Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

Kalm, K., & Norris, D. (2014). The Representation of Order Information in Auditory-Verbal Short-Term Memory. *Journal of Neuroscience*, *34*(20), 6879-6886. https://doi.org/10.1523/JNEUROSCI.4104-13.2014

Kanerva, P. (1988). *Sparse Distributed Memory*.

Kowialiewski, B., Gorin, S., & Majerus, S. (2021). Semantic knowledge constrains the processing of serial order information in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001031

Kowialiewski, B., Lemaire, B., & Portrat, S. (2021). How does semantic knowledge impact working memory maintenance? Computational and behavioral investigations. *Journal of Memory and Language*, *117*(104208). https://doi.org/10.1016/j.jml.2020.104208

Kowialiewski, B., & Majerus, S. (2018). The non-strategic nature of linguistic long-term memory effects in verbal short-term memory. *Journal of Memory and Language*, *101*, 64-83. https://doi.org/10.1016/j.jml.2018.03.005

Kowialiewski, B., & Majerus, S. (2020). The varying nature of semantic effects in working memory. *Cognition*, *202*, 104278. https://doi.org/10.1016/j.cognition.2020.104278

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211.

Lewandowsky, S., & Farrell, S. (2008). Short-Term Memory : New Data and a Model. In *Psychology of Learning and Motivation* (Vol. 49, p. 1-48). Elsevier. https://doi.org/10.1016/S0079-7421(08)00001-7

Logie, R. H., Saito, S., Morita, A., Varma, S., & Norris, D. (2016). Recalling visual serial order for verbal sequences. *Memory & Cognition*, *44*(4), 590-607. https://doi.org/10.3758/s13421-015-0580-9

Majerus, S. (2013). Language repetition and short-term memory : An integrative framework. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00357

Majerus, S., D'Argembeau, A., Martinez Perez, T., Belayachi, S., Van der Linden, M., Collette, F., Salmon, E., Seurinck, R., Fias, W., & Maquet, P. (2010). The Commonality of Neural Networks for Verbal and Visual Short-term Memory. *Journal of Cognitive Neuroscience*, *22*(11), 2570-2593. https://doi.org/10.1162/jocn.2009.21378

Mensink, G.-J., & Raaijmakers, J. (1988). A Model for Interference and Forgetting. *Psychological Review*, *95*(4), 434-455.

Monnier, C., & Bonthoux, F. (2011). The semantic-similarity effect in children : Influence of long-term knowledge on verbal short-term memory: Semantic contribution to serial recall in children. *British Journal of Developmental Psychology*, *29*(4), 929-941. https://doi.org/10.1111/j.2044-835X.2010.02024.x

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169-177. https://doi.org/10.3758/s13428-012-0243-8

Morey, R. D., & Rouder, J. N. (2014). *BayesFactor : Computation of Bayes factors for common designs (Version 0.9.2+)*.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251-269. https://doi.org/10.3758/BF03213879

Nairne, J. S., & Kelley, M. R. (2004). Separating item and order information through process dissociation. *Journal of Memory and Language*, *50*(2), 113-133. https://doi.org/10.1016/j.jml.2003.09.005

Neale, K., & Tehan, G. (2007). Age and redintegration in immediate memory and their relationship to task difficulty. *Memory & Cognition*, *35*(8), 1940-1953. https://doi.org/10.3758/BF03192927

Neath, I. (1997). Modality, concreteness, and set-size effects in a free reconstruction of order task. *Memory & Cognition*, *25*(2), 256-263. https://doi.org/10.3758/BF03201116

Nee, D. E., & Jonides, J. (2013). Neural evidence for a 3-state model of visual short-term memory. *NeuroImage*, *74*, 1-11. https://doi.org/10.1016/j.neuroimage.2013.02.019

Nimmo, L. M., & Roodenrys, S. (2004). Investigating the phonological similarity effect : Syllable structure and the position of common phonemes. *Journal of Memory and Language*, *50*(3), 245-258. https://doi.org/10.1016/j.jml.2003.11.001

Oberauer, K. (2002). Access to information in working memory : Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411-421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K. (2009). Design for a Working Memory. In *Psychology of Learning and Motivation* (Vol. 51, p. 45-100). Elsevier. https://doi.org/10.1016/S0079-7421(09)51002-X

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory : A computational implementation of the Time-Based Resource-Sharing theory. *Psychonomic Bulletin & Review*, *18*(1), 10-45. https://doi.org/10.3758/s13423-010-0020-6

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory : An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779-819. https://doi.org/10.3758/s13423-012-0272-4

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21-59. https://doi.org/10.1037/rev0000044

Osgood, C. E. (1949). The similarity paradox in human learning : A resolution. *Psychological Review*, *56*(3), 132-143. https://doi.org/10.1037/h0057488

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. https://doi.org/10.1080/14640749508401396

Poirier, M., Yearsley, J. M., Saint-Aubin, J., Fortin, C., Gallant, G., & Guitard, D. (2019). Dissociating visuo-spatial and verbal working memory : It’s all in the features. *Memory & Cognition*, *47*(4), 603-618. https://doi.org/10.3758/s13421-018-0882-9

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509-522. https://doi.org/10.1037/0278-7393.2.5.509

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*(2), 270-279. https://doi.org/10.3758/s13414-013-0605-z

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

Saint-Aubin, J., & Poirier, M. (1999a). The Influence of Long-term Memory Factors on Immediate Serial Recall : An Item and Order Analysis. *International Journal of Psychology*, *34*(5-6), 347-352.

Saint-Aubin, J., & Poirier, M. (1999b). 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. https://doi.org/10.1080/713755814

Saint-Aubin, J., Yearsley, J. M., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term : The cost of relative distinctiveness. *Journal of Memory and Language*, *118*, 104219. https://doi.org/10.1016/j.jml.2021.104219

Saito, S., Logie, R. H., Morita, A., & Law, A. (2008). Visual and phonological similarity effects in verbal immediate serial recall : A test with kanji materials. *Journal of Memory and Language*, *59*(1), 1-17. https://doi.org/10.1016/j.jml.2008.01.004

Schneegans, S., & Bays, P. M. (2017). Neural Architecture for Feature Binding in Visual Working Memory. *The Journal of Neuroscience*, *37*(14), 3913-3925. https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, *21*(2), 168-175. https://doi.org/10.3758/BF03202729

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. https://doi.org/10.1080/17470218.2011.604787

Tyler, L. K., Moss, H. E., Galpin, A., & Voice, J. K. (2002). Activating meaning in time : The role of imageability and form-class. *Language and Cognitive Processes*, *17*(5), 471-502. https://doi.org/10.1080/01690960143000290

Visscher, K. M., Kaplan, E., Kahana, M. J., & Sekuler, R. (2007). Auditory Short-Term Memory Behaves Like Visual Short-Term Memory. *PLoS Biology*, *5*(3), e56. https://doi.org/10.1371/journal.pbio.0050056

Watkins, M. J., & Watkins, O. C. (1976). Cue-overload theory and the method of interpolated attributes. *Bulletin of the Psychonomic Society*, *7*(3), 289-291. https://doi.org/10.3758/BF03337192

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences : Comparing phonological similarity and pitch proximity. *Memory & Cognition*, *38*(2), 163-175. https://doi.org/10.3758/MC.38.2.163