Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2023 American Psychological Association
ISSN: 0278-7393

https://doi.org/10.1037/xlm0001279

Does Semantic Similarity Affect Immediate Memory for Order? Usually Not, But Sometimes It Does

Benjamin Kowialiewski[1, 2], Steve Majerus[2, 3], and Klaus Oberauer[1]
[1] Department of Psychology, University of Zürich
[2] Psychology & Neuroscience of Cognition Research Unit, University of Liège
[3] Fund for Scientific Research FNRS, Brussels, Belgium

Recall performance in working memory (WM) is strongly affected by the similarity between items. When asked to encode and recall list of items in their serial order, people confuse more often the position of similar compared to dissimilar items. Models of WM explain this deleterious effect of similarity through a problem of discriminability between WM representations. In contrast, when lists of items that are all semantically similar are compared to lists of dissimilar items, semantic similarity does not negatively impact order memory, questioning the idea that semantic information is part of the WM content. This study reports four experiments in which semantic similarity was manipulated using lists composed of multiple semantic categories. These experiments revealed two main patterns. First, semantic similarity can *increase*, rather than decrease, order memory. Second, semantic knowledge reliably constrains the way items migrate; when migrating, items tend to do so more often toward the position of other similar items, rather than migrating toward other dissimilar items. These results challenge the way current models of WM represent similarity. The plausibility of different theoretical accounts and mechanisms is discussed.

Keywords: working memory, semantic similarity, order memory, semantic knowledge

Working memory (WM) performance is affected by the similarity between items. People tend to confuse similar items in their order more often than dissimilar ones, a classical phenomenon observed across a variety of domains (Guitard & Cowan, 2020; Gupta et al., 2005; Jalbert et al., 2008; Lin & Luck, 2009; Logie et al., 2016; Saito et al., 2008; Visscher et al., 2007). In contrast to other types of similarity, semantic similarity does not decrease people's ability to recall items in their presentation order (Kowialiewski, Krasnoff, et al., 2023; Neale & Tehan, 2007; Poirier & Saint-Aubin, 1995; Saint-Aubin & Poirier, 1999). Still, there are some studies—reviewed below—showing subtle but intriguing effects of semantic similarity on order memory. The purpose of the present study is to determine under which conditions such effects are found, and how they come about.

Verbal WM is often tested through serial recall. In this task, participants are asked to encode lists of sequentially presented items and recall them in their original presentation order. Serial recall involves two different demands. First, people need to maintain the *identity* of the to-be-remembered items. Many errors in serial recall are failures to retrieve items' identity, such that participants fail to report an item at all (i.e., omission errors) or produce items that were not part of the list (i.e., extra-list intrusions). For instance, given the sequence "ABCDEF," people often output sequences such as "A*DCZF" (where "*" and "Z" refer to an omission and an extra-list intrusion, respectively). Second, serial recall involves reporting items at their correct serial positions. Participants often produce *transposition errors*, which refer to items migrating toward wrong serial positions. In a sequence such as "ABCDEF," people often output responses such as "ABDCEF." Evidence supports independent processes subtending the maintenance of item identity and order information (Majerus, 2013, 2019; Nairne & Kelley, 2004). Item and order memory are differentially affected by dual-task interference (Gorin et al., 2016; Henson et al., 2003). The processing of item and order information also recruits distinct neural substrates, as shown by neuroimaging, direct electrical stimulation, and neuropsychological data (Kalm & Norris, 2014; Majerus, 2013; Majerus et al., 2010, 2015; Papagno et al., 2017).

At a theoretical level, the ability to recall the identity of items is assumed to reflect the temporary activation of long-term memory knowledge, as proposed by an embedded processes view of WM (Cowan, 1999; Oberauer, 2002, 2009). In the case of verbal stimuli, this temporary activation is assumed to occur in the linguistic system (Majerus, 2013, 2019; Martin & Saffran, 1997). This activation of long-term memory representations is, however, not sufficient to maintain the order of items in a list, because the sequential arrangement of items is typically new and arbitrary (Norris, 2017, 2019). Therefore, items' order needs to be maintained in some way. One proposed mechanism is the creation of temporary item–context associations (Burgess & Hitch, 1999, 2006; Lewandowsky, 1999; Lewandowsky & Farrell, 2008; Oberauer et al., 2012; Schneegans & Bays, 2017). In serial recall, the nature of the context is positional. For instance, when encoding the sequence "monkey, ball, desk," the linguistic features of "monkey" and "ball" are associated to "position 1" and "position 2,"

Benjamin Kowialiewski https://orcid.org/0000-0002-7743-761X

This research was supported by a grant from the Swiss National Science Foundation to Klaus Oberauer (project 100014_192204).

All the data and codes have been made available on the Open Science Framework: https://osf.io/wzndt/ (Kowialiewski, Majerus, & Oberauer, 2023).

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

respectively. Retrieval is then performed by sequentially reactivating the positional markers one by one (e.g., starting by cueing "monkey" with "position 1"). Current evidence suggests this item–context binding as a plausible mechanism for maintaining items' order. For instance, models implementing this mechanism predict the pattern of order errors observed in serial recall: People are more likely to transpose items presented at adjacent versus distant serial positions, a phenomenon also called *the locality constraint* (Henson, 1998). The locality constraint is a consequence of the property of positional markers to which items are associated to: Adjacent positional markers are assumed to be more similar than distant ones. When items are cued, other items sharing these markers are also partially retrieved, thus increasing the probability to retrieve an adjacent than distant item when a transposition occurs. The binding mechanism also predicts that it should be possible to retrieve a context when an item is given (i.e., retrieving "position 1" from the word "monkey"), a prediction which has received empirical support in the verbal WM literature (Guérard et al., 2009; Kowialiewski, Krasnoff, et al., 2023).

Similarity and Order Memory

The similarity between items negatively impacts order memory. The most typical effect is the *phonological similarity* effect (Camos et al., 2013; Fallon et al., 2005; Gupta et al., 2005; Karlsen et al., 2007; Lian & Karlsen, 2004; L. Nimmo & Roodenrys, 2005; L. M. Nimmo & Roodenrys, 2004; Roodenrys, Guitard, et al., 2022), in which transposition errors increase for phonologically similar (e.g., cat, bat, fat, mat, rat) compared to phonologically dissimilar (e.g., wall, desk, car, dig, arm) items. Similarity-based confusions have been observed across a wide variety of 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) domains, suggesting that similarity-based confusions reflect a general property of WM. At a theoretical level, similarity effects are explained by a discriminability problem between WM representations. When trying to retrieve the word "cat" from "position 1," this likely leads to the retrieval of a degraded representation of the original trace (i.e., retrieving "_at") which needs to be compared to items stored in long-term memory (Oberauer et al., 2012; Schweickert, 1993). If other list items are similar to the target (e.g., "mat" and "fat"), it is more likely to select these competitors instead of the original target item than if the other list items are dissimilar (e.g., "wall" and "desk"). Similarity-based confusions have implications for our theories of WM, as they indicate what kind of representation is bound to context.

Contrary to other types of similarity, semantic similarity does not increase transposition errors. Semantic similarity is classically manipulated by comparing WM performance for pure lists of semantically similar (e.g., cheetah, puma, lion, panther, lynx, tiger) and dissimilar (e.g., liver, mallet, wasp, mug, cushion, taxi) words. When participants are asked to memorize and recall lists of words in their presentation order, memory performance increases for lists composed of semantically similar as opposed to dissimilar words (Kowialiewski & Majerus, 2020; Nairne & Kelley, 2004; Neale & Tehan, 2007; Poirier & Saint-Aubin, 1995; Saint-Aubin & Poirier, 1999; Tse, 2009, 2010; Tse et al., 2011). This recall advantage is characterized by an increase of item memory. In contrast, semantic similarity does not impair order memory (Neale & Tehan, 2007; Poirier & Saint-Aubin, 1995; Saint-Aubin & Poirier, 1999).

Some studies found a deleterious impact of semantic similarity on memory for order (Saint-Aubin et al., 2005; Tse, 2010; Tse et al., 2011), and some authors have argued that this might be explained by the way semantic similarity is manipulated (Ishiguro & Saito, 2020). However, several recent studies using large samples have shown a systematic absence of detrimental effect of semantic similarity on memory for order across a variety of experimental conditions and different semantic similarity metrics (Kowialiewski, Krasnoff, et al., 2023; Neath et al., 2022).[1] If WM encoded semantic information through item-context binding, we expect that semantically similar items should be recalled more often in the wrong order than semantically dissimilar items. This is because items' semantic features should be more difficult to discriminate in semantically similar lists. Therefore, current evidence indicates that semantic features are likely not encoded in WM through item-context binding.

Cases of Effects of Semantic Similarity on Order Memory

The null effect of semantic similarity on order memory is a well-replicated phenomenon (Kowialiewski, Krasnoff, et al., 2023; Neale & Tehan, 2007; Neath et al., 2022; Poirier & Saint-Aubin, 1995, 1996; Saint-Aubin & Poirier, 1999). However, most manipulations of semantic similarity involved lists composed of purely similar or dissimilar items. More fine-grained manipulations of semantic similarity reveal that semantic similarity, although not decreasing order memory, can constrain the pattern of serial order errors.

Poirier et al. (2015) manipulated semantic similarity by presenting lists in which the three first items were semantically similar (e.g., officer, badge, siren, music, tourist, yellow). They compared this control condition to an experimental condition in which the fifth item was semantically similar to the three first items (e.g., officer, badge, siren, fence, police, tractor). They observed that the fifth item tended to be transposed more often toward positions 1, 2, and 3 in the experimental as compared to the control condition. Equivalent results have recently been reported by Kowialiewski, Gorin, and Majerus (2021). They manipulated semantic similarity by presenting two categories composed of three items. In one condition, similar items were presented in a grouped fashion (e.g., leaf, tree, branch, *cloud*, *sky*, *rain*). In another condition, similar items were presented in an interleaved fashion (e.g., leaf, *cloud*, tree, *sky*, branch, *rain*). Kowialiewski, Gorin, and Majerus (2021) observed that the semantically similar items, when migrating, tended to do so toward the position of other semantically similar items (*within-category transpositions*), rather than toward the position of semantically dissimilar items (*between-category transpositions*), compared to the same positions in the semantically dissimilar condition. In other words, when items migrated, they tended to do so more often to the positions of other semantically similar items. This result was however observed only when the items were grouped. When items were presented in an interleaved fashion, no increased within-category transpositions were observed.

In addition to the above observations, a recent study showed that semantic similarity can *increase* the ability to recall items in their

[1] Note that Ishiguro and Saito (2020) proposed that semantic similarity might be better characterized along a three-dimensional space encompassing, valence, arousal, and dominance (Moors et al., 2013). The recent study by Kowialiewski, Krasnoff, et al., 2023 reported a regression analysis showing a systematic failure to show an effect of this metric on order memory.

correct order. Kowialiewski et al. (2022) reported a series of experiments in which lists were presented in subgroups of two semantically similar items (e.g., **car**, **taxi**, shirt, vest, *orange*, *lemon*). As compared to a dissimilar condition, they observed that semantically similar subgroups not only enhanced item memory, but also order memory. This result is surprising, because previous semantic similarity manipulations using pure lists systematically failed to produce any impact on overall order memory, as reviewed above.

There remain empirical uncertainties regarding the specific conditions under which semantic similarity does impact order memory. Poirier et al. (2015) reported migration errors occurring across the entire list (i.e., item five migrating toward positions 1, 2, and 3). The results reported by Kowialiewski, Gorin, and Majerus (2021) suggest, however, that these migration errors might be more limited and local, as they were found only when similar items were presented in groups, but not when interleaved. Both studies have methodological limitations. In the Poirier et al. study, transposition errors were not corrected by the total number of order errors, and not even by the total number of items recalled. Instead, these authors used the proportion of trials as dependent variable. As people remember more similar than dissimilar items, they are expected to recall more of these items in the correct order *and* in the incorrect order in absolute term, even if the probability to recall each item in its correct position is equivalent. We report in Appendix A a reanalysis of their data[2] in which transposition errors were corrected by the total number of order errors. The results of this analysis indicate that their results were robust to the change of dependent variable. The experiment reported by Kowialiewski, Gorin, and Majerus (2021) may have suffered from a lack of statistical power. Each experimental condition involved only 15 trials, which may be insufficient to properly compute within-group transpositions (i.e., see statistical procedure below). The null effect found in their study might therefore be a false negative.

The Present Study

To sum up, semantic similarity does not negatively impact order memory, contrary to other types of similarity. This raises the possibility that semantic information is not bound to contexts. At the same time, a few recent studies reported that semantic similarity constrains the pattern of transposition errors in WM, and also increases order memory. If semantic is not bound to context, it is difficult to explain why it does influence order memory in some cases. Current evidence for an effect of semantic knowledge on order memory remains however scarce and not well specified. Therefore, before considering potential effects of semantic knowledge of order memory in our models of WM, we need stronger empirical evidence supporting it.

This study investigates the boundary conditions under which semantic knowledge influences memory for serial order in WM. We tested the impact of semantic knowledge not only on overall order memory, but also on the pattern of within-category transpositions. Experiment 1 is a conceptual replication of Kowialiewski, Gorin, and Majerus (2021) results using a full within-subject design and an immediate serial recall task. Experiment 2 was equivalent to Experiment 1, except that we used an order reconstruction task, which maximally encourages the encoding of order information. Experiment 3 directly addressed the impact of semantic similarity on overall order memory performance using a between-subject design. Finally, Experiment 4 tested whether the increased within-category transposition phenomenon is a strategic versus nonstrategic process.

Experiment 1

In Experiment 1, we manipulated semantic similarity across three conditions. In one condition, items drawn from two different categories were presented in a grouped manner (e.g., leaf, tree, branch, *sky*, *cloud*, *rain*). In another condition, items drawn from two different categories were presented in an interleaved manner (e.g., leaf, *sky*, tree, *cloud*, branch, *rain*). In a control condition, all items were drawn from distinct semantic categories (e.g., dog, wall, planet, arm, grass, key). Contrary to the study conducted by Kowialiewski, Gorin, and Majerus (2021), the present experiment manipulated the grouped and interleaved condition with a full within-subject design to allow a direct comparison between all conditions.

Method

Transparency and Openness

All the data, codes, and materials across all experiments have been made available on Open Science Framework (OSF): https://osf.io/wzndt/. This study's design and its analysis were not preregistered.

Participants

We recruited 32 participants from the university community of the University of Liège. Here and in the following experiments, the sample size was chosen because it has been sufficient for obtaining strong evidence for similarity effects in previous experiments and was feasible with the available resources. As we planned to analyze the data with Bayesian statistics, this way of determining the sample size does not involve a risk of biasing the outcome: If the sample size had been too small, this would have been reflected in ambiguous Bayes factors (BFs)—a limitation that could be overcome by increasing the sample size (Rouder, 2014).

All participants were French native speakers, reported no history of neurological disorder or learning difficulties. All participants gave their electronic informed consent before starting the experiment. The experiment had been approved by the Ethic committee of the Faculty of Psychology, Speech and Language Therapy, and Education of the University of Liège.

Material

The stimuli used in Experiment 1 were French nouns. The stimuli were drawn from 40 different taxonomic categories. There were three items per category, resulting in a total of 120 items included in the full set of stimuli. Examples of categories involved body parts (arm, thigh, leg), vehicles (bus, car, truck), or animals (tiger, cheetah, panther). The full list of stimuli has been made available on OSF, along with an English translation. The semantically similar lists were built by using items from two different categories. The semantically grouped lists were created first. The semantically interleaved lists were created using the same sequences as the semantically grouped sequences but changing the items' order such that the semantically similar items were now presented in an interleaved fashion. The semantically dissimilar lists were built by randomly sampling items from six different categories. Each item was used

[2] We are grateful to Marie Poirier and Jean Saint-Aubin for sharing their data with us.

three times across the whole experiment: once in a grouped condition, once in an interleaved condition, and once in a dissimilar condition. This way of creating the experimental list led to 20 trials per condition. Each item was recorded by a native male French-speaking individual in a neutral voice in a soundproof booth. A noise reduction filter was applied to all the stimuli to suppress the residual background noise using the Audacity software. Items were then isolated on separate audio files.

Procedure

Due to the COVID-19 pandemic, all participants were tested remotely. The experimenters met the subjects via the LifeSize® desktop application and gave instructions. The participants performed the experiment separately on an online webpage coded in HTML5 and JavaScript. The auditory presentation of the stimuli was made possible using the Pizzicato library (https://alemangui.github.io/pizzicato/). Participants' oral responses were recorded via the LifeSize application for later transcription. All responses were transcribed by an internship student.

Each trial started with the presentation of a six-item list to be remembered. Each item was auditorily presented at a pace of 1,000 ms per item. A pilot experiment indicated that this presentation rate resulted in reasonable recall performance levels (~70%). Directly after the presentation of the last to-be-remembered item, participants were asked to recall out loud the items in their original presentation order. If participants could not recall an item, they were invited to say the word "blanc" (i.e., "blank" in French), resulting in an omission error. After recalling all items, participants were required to press the spacebar of their keyboard to initiate the next trial. Before the beginning of the main experiment, participants performed three training trials. During the training trials, compliance with task instructions was checked by the experimenter. Between trials, participants had the opportunity to adjust the level of their speaker or headphones if they needed to.

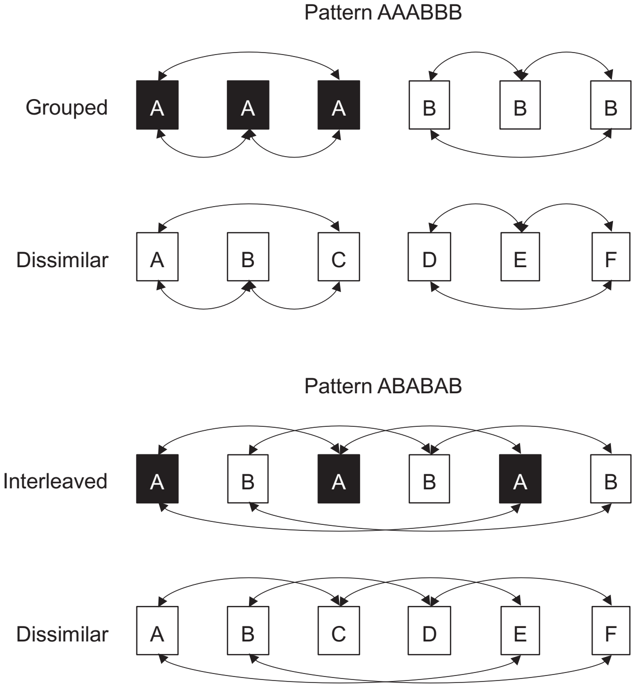
Scoring Procedure

Recall performance was assessed using two different scoring procedures. First, we used an item recall scoring procedure, in which an item was considered correct if recalled at all, regardless of the serial position at which it was recalled. For instance, given the target sequence "ABCDEF" and the recalled sequence "A*BD*E" (where the symbol "*" represents an omission), items "A," "B," "D," and "E" would be considered correct. This criterion gives a good measure of the ability to recall item information, without considering serial order. Second, we used a conditionalized order recall score, in which an item, when recalled, was scored correct if recalled in the correct serial position. As items not recalled at all are not diagnostic of order memory, they were scored as missing data. In the previous example, the sequence would be scored as [1, 0, NA, 1, 0, NA], thus resulting in an average order recall score of 2/4. This criterion gives a measure of order memory which is independent of the influence of item memory. As people are expected to recall more semantically similar compared to dissimilar items, the number of items recalled must be accounted for.

Next, we quantified which proportion of transposition errors were within-category transpositions. To this end, we counted the number of within-category and of between-category transpositions in each condition.

A within-category transposition is defined as a transposition occurring between two semantically similar items. For the *grouped* condition following the AAABBB pattern as displayed in Figure 1, upper panel, these transpositions involved positions [1, 2, 3] or positions [4, 5, 6]. For instance, given the sequence "ABCDEF," recalling "ACBDEF" constitutes two within-category transposition. In contrast, recalling "ABDCEF" constitutes two between-category transpositions, because two items migrated across group boundaries in the AAABBB pattern. For the *interleaved* condition, within-category transposition errors were counted according to the ABABAB pattern as displayed in Figure 1, lower panel, involved transpositions occurring between positions [1, 3, 5] or positions [2, 4, 6]. For instance, given the sequence "ABCDEF," recalling "ADCBEF" constitutes two within-category transpositions. In contrast, recalling "BACDEF" constitutes two between-category transpositions, because two items migrated into positions occupied by the other semantic group within the ABABAB pattern. After classifying these transposition errors, the total number of within-category transpositions in each experimental condition was divided by the total number of transposition errors occurring in the same condition. For instance, if participants produced a total of 12 within-category transpositions and five between-category transpositions in an experimental condition, the proportion of within-category transpositions for this condition was $12/(12 + 5) = 0.706$. Similarly, if participants produced five within-category transpositions and three between-category transpositions, the proportion of within-category

Figure 1
Patterns of Transposition Errors Used Across All Experiments



Note. Pattern AAABBB: Transposition errors occurring between items in positions [1, 2, 3] or in positions [4, 5, 6] are counted as within-category transpositions. Pattern ABABAB: Transposition errors occurring between items in positions [1, 3, 5] or in positions [2, 4, 6] are counted as within-category transpositions.

transpositions for this condition was 5/(5 + 3) = 0.625. For the dissimilar condition, which served as a control condition for the grouped and interleaved condition, we analyzed transposition errors twice, once using each of the two patterns in Figure 1. As in this condition, there are no two words from the same category, the pseudo within-category transpositions (i.e., transpositions within each subset of the applied pattern) were treated as controls for the condition in which semantic categories were distributed according to that pattern.

This analysis tests the impact of a given semantic structure on the *pattern* of transposition errors, not the transposition rate itself. A high within-category proportion does *not* mean that more order errors were produced in total. Instead, it means that the experimental condition elicited a stronger constraint on serial order errors. In other words, a high within-category proportion implies that when a transposition occurred, it did so most of the time following one of the patterns displayed in Figure 1. The dissimilar condition served as a control to test what would happen if patterns of transposition errors were not constrained by a semantic structure.

Statistical Analysis

We conducted Bayesian analyses using the BF package implemented in R. We use the classification of strength of evidence proposed by Jeffreys (1998): a BF of 1 provides no evidence, $1 < \text{BF} < 3$ provides anecdotal evidence, $3 < \text{BF} < 10$ provides moderate evidence, $10 < \text{BF} < 30$ provides strong evidence, $30 < \text{BF} < 100$ provides very strong evidence, and $100 < \text{BF}$ provides extreme/decisive evidence. In Bayesian analysis of variances (ANOVAs), we performed Bayesian model comparisons using a top-down testing procedure. We first started with the most complex possible structure including all effects, their interactions, a random intercept, and the random slopes for each main effect. We then progressively reduced the model's complexity by comparing the current best model to the same model without the specific effect of interest. We first started by testing the random slopes associated with each main effect, followed by the interactions and the main fixed effects until reaching the best possible model. We assessed each effect of interest by comparing the best model to the same model with or without the effect in question. To minimize the error of model estimation, the number of Monte Carlo simulations generated was set to $N_{\text{iterations}} = 10^4$. We used the default Cauchy prior distribution with a medium scale, $r = \frac{\sqrt{2}}{2}$. In addition, each graph reports the 95% within-subject confidence intervals for each mean, following the recommendations made by Baguley (2012).

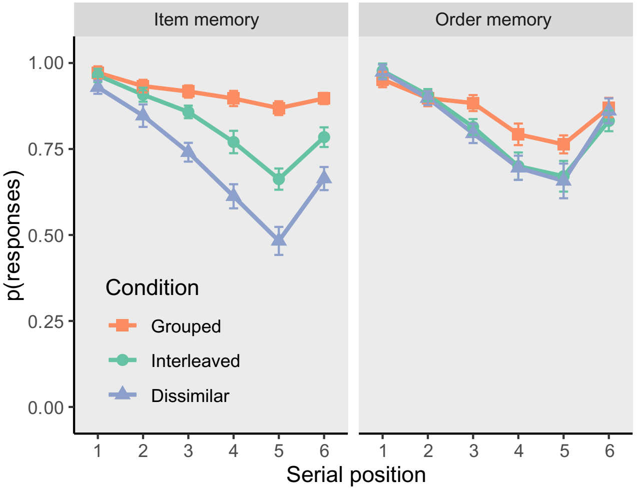
For simplicity, we tested only the effects of interest. We report in Appendix B detailed BFs for all experiments.

Results

Recall performance as a function of semantic condition (grouped, interleaved, dissimilar) and serial position (one through six) was assessed using Bayesian repeated-measures ANOVAs. First, recall performance was analyzed using the item recall criterion as dependent variable. The best model was the full model. Compared to the best model, we found decisive evidence supporting an effect of semantic similarity ($\text{BF}_{10} = 2.153\text{e}{+}21$). Bayesian paired-samples *t* tests showed that both the grouped ($\text{BF}_{10} = 7.847\text{e}{+}12$) and the interleaved ($\text{BF}_{10} = 1.522\text{e}{+}8$) conditions were better recalled than the dissimilar condition. The grouped condition was also better recalled than the interleaved condition, as supported by decisive evidence ($\text{BF}_{10} = 1.772\text{e}{+}8$). These results are illustrated in Figure 2, left panel.

Figure 2

Experiment 1—Serial Position Curves for Recall Performance



Note. Recall performance as a function of semantic condition (grouped, interleaved, dissimilar) and serial position (one through six). Left panel: Item recall criterion. Right panel: Order recall criterion. Error bars represent 95% within-subject confidence intervals. See the online article for the color version of this figure.

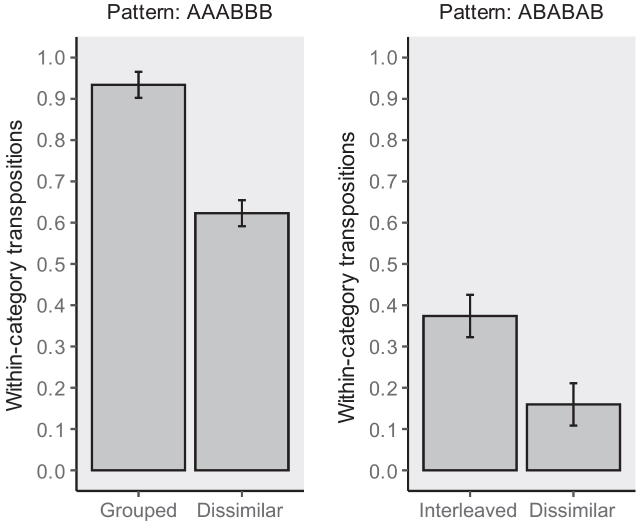
Next, recall performance was analyzed using the order recall criterion as dependent variable. The best model was the full model. Compared to the best model, we found moderate evidence supporting an effect of semantic similarity ($\text{BF}_{10} = 4.031$). Bayesian paired-samples *t* tests showed that the grouped condition led to better order recall than the interleaved ($\text{BF}_{10} = 107.311$) and dissimilar ($\text{BF}_{10} = 151.729$) conditions. No credible difference between the interleaved and dissimilar conditions emerged ($\text{BF}_{10} = 0.521$). These results are illustrated in Figure 2, right panel.

The next analysis assessed the impact of semantic similarity on patterns of transposition errors. These patterns were analyzed using Bayesian paired-samples *t* tests. In the AAABBB pattern, we found an increased proportion of within-category transpositions in the grouped as compared to the dissimilar condition, and this difference was supported by decisive evidence ($\text{BF}_{10} = 2.491\text{e}{+}8$). In the ABABAB pattern, we also found an increased proportion of within-category transpositions in the interleaved as compared to the dissimilar condition, and this difference was supported by decisive evidence ($\text{BF}_{10} = 149$). Numerical comparison of effect sizes indicates that the magnitude of the effect was twice as big in the grouped ($d = 1.81$) compared to the interleaved ($d = 0.766$) conditions.[3] To sum up these results, semantic similarity constrained the pattern of transposition errors across both the grouped and interleaved conditions, as illustrated in Figure 3.

[3] Note that since the two patterns of within-category transpositions have very different proportions in the dissimilar condition, the interpretation of a potential interaction would be hazardous. The magnitude of both patterns was therefore not tested.

Figure 3

Experiment 1—Increase of Within-Category Transposition Proportions as a Function of List Structure and Semantic Condition



Note. Proportion of within-transpositions among all transpositions across semantic conditions. Left panel: Pattern AAABBB, grouped versus dissimilar condition. Right panel: Pattern ABABAB, interleaved versus dissimilar condition. Error bars represent 95% within-subject confidence intervals.

Discussion

As has been observed many times, semantic similarity enhanced item memory. Item memory was better in the grouped than in the interleaved condition, despite these lists being identical in terms of semantic content, differing only in their order of presentation. This *separation effect* replicates previous observations (Kowialiewski, Gorin, & Majerus, 2021; Kowialiewski et al., 2022; Saint-Aubin et al., 2014).

Critically, we also observed an impact of semantic similarity on order memory. First, order memory was increased when items were semantically grouped but not when semantic similarity was manipulated in an interleaved fashion. Second, semantic similarity constrained transposition in both semantic conditions. When an item migrated, it did so more often toward the position of another similar item, compared to yoked positions in the dissimilar condition. These results replicate those already observed by Kowialiewski, Gorin, and Majerus (2021) and extend them by showing an increased proportion of within-category transposition errors also for the interleaved condition. It is possible that the earlier study failed to uncover this effect due to a smaller number of trials, resulting in reduced statistical power.

Experiment 2

The purpose of Experiment 2 was to replicate the results of Experiment 1 by using an order reconstruction task, in which items are given at retrieval, and participants are asked to reconstruct their presentation order. As all list items are present at retrieval, item errors are not possible, providing a direct measure of participants' ability to maintain the serial order of items.[4] If the results observed in Experiment 1 are robust, they should replicate when switching from serial recall to order reconstruction. We changed the set size from six to eight items to increase the occurrence of order errors. With this experiment, we maximized the possibility to observe any effect that would be otherwise difficult to detect. We manipulated semantic similarity as in Experiment 1, by using two categories of similar items presented in a grouped and interleaved fashion. These conditions were compared to dissimilar condition. We expected to replicate the results from Experiment 1, that is, increased order memory when items are semantically grouped, and higher within-category transposition proportions in the grouped and interleaved conditions.

Method

Participants

Thirty-two adults aged between 18 and 35 participated in this experiment. The sample size was chosen based on Experiment 1. 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

The procedure used to create the lists matched the one used in Experiment 1, with a few exceptions. First, as we increased set size from six to eight, all stimuli were drawn from 40 semantic categories composed of four items each. Second, as the experiment was conducted on English native speakers, all stimuli were now English (as opposed to French) nouns. Third, we switched the presentation modality from auditory to visual because of the constraints imposed by the order reconstruction task (i.e., at test, items must be displayed on screen). The list of stimuli is available on OSF.

Procedure

Participants were asked to encode eight-word lists. Each word was sequentially presented at a pace of one item/s (1,000 ms ON, 0 ms OFF) and appeared at the center of the screen in Courier font in lowercase. After presentation of the last item, all memoranda appeared again on screen in two rows in uppercase letters. The words were arranged in a pseudorandom order, which for each trial consisted of random sampling without replacement from all possible permutations of the sequence [1:8], except the first permutation, which is the original presentation order. Participants clicked on each item in the order in which they originally appeared. Words were replaced by a string of "#" characters matching the original word's length after each click. This latter constraint explicitly prevents repetitions.

[4] One could argue that item errors are still possible in order reconstruction. There is evidence showing that individual items' properties, such as their concreteness, can impact performance in order reconstruction (Neath, 1997). The fact that items' properties sometimes influence order reconstruction does n't mean that item errors are possible in this paradigm. Instead, it means that order errors can be influenced by variables such as concreteness or word length. This is to be expected whenever item properties influence the strength or precision with which an item is bound to its positional context.

Participants performed three training trials before beginning the main experiment.

Scoring Procedure

In order reconstruction, item memory is always perfect, as all items are available at retrieval. Therefore, we only computed an order memory score: An item was considered correct if retrieved at the correct serial position. Patterns of transposition errors were analyzed the same way as in Experiment 1.

Statistical Analysis

Statistical analyses were identical to Experiment 1.

Results

The proportion of correct response as a function of semantic similarity (grouped, interleaved, dissimilar) and serial position (one through eight) was assessed using a Bayesian repeated-measures ANOVA. The best model was the full model. Comparing the best model to one removing the fixed effect of semantic similarity, we found decisive evidence supporting that effect ($BF_{10} = 4.653e+4$). Bayesian paired-samples *t* tests show that the proportion of correct responses was higher in the grouped compared to the dissimilar condition, and this difference was supported by decisive evidence ($BF_{10} = 3{,}390$). Similarly, there was a recall advantage in the grouped over the interleaved condition, supported by very strong evidence ($BF_{10} = 71.69$). There was a slight difference between the interleaved and dissimilar conditions, but this difference was only supported by anecdotal evidence ($BF_{10} = 2.47$). These results, also illustrated in Figure 4, can be summarized as follow: grouped > interleaved ~ dissimilar.

Next, we performed Bayesian paired-samples *t* tests on the within-category transposition proportions. In Pattern AAABBB, we found increased within-category transposition proportions in the grouped compared to the dissimilar condition, supported by decisive evidence ($BF_{10} = 1.840e+4$). In pattern ABABAB, we found the same effect, supported by strong evidence ($BF_{10} = 13$). The magnitude of this effect was numerically larger in the grouped ($d = 1.077$) than in the interleaved ($d = 0.573$) condition. These results are displayed in Figure 5.

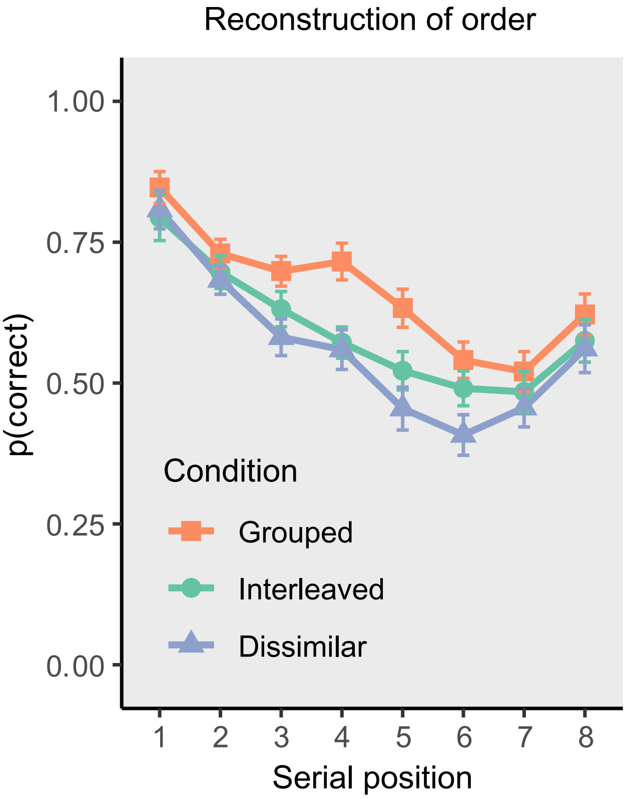
Discussion

Replicating Experiment 1, only the grouped condition credibly enhanced order memory. Both the grouped and interleaved semantic conditions consistently constrained the pattern of serial order errors. This means that the absence of a semantic similarity effect in the interleaved condition observed by Kowialiewski, Gorin, and Majerus (2021) may have been caused by a lack of power.

The fact that order recall benefits from semantic grouping is in striking contrast with the usual null effect of semantic similarity on order memory, when pure lists of semantically similar or dissimilar words are compared (Kowialiewski, Krasnoff, et al., 2023; Neale & Tehan, 2007; Neath et al., 2022; Poirier & Saint-Aubin, 1995; Saint-Aubin & Poirier, 1999). However, as we did not include pure semantically similar lists, we cannot rule out the possibility that there are some peculiarities in our materials by which semantic similarity improved order memory even without a grouping structure. The next experiment tested this possibility.

Figure 4

Experiment 2—Serial Position Curves for Recall Performance



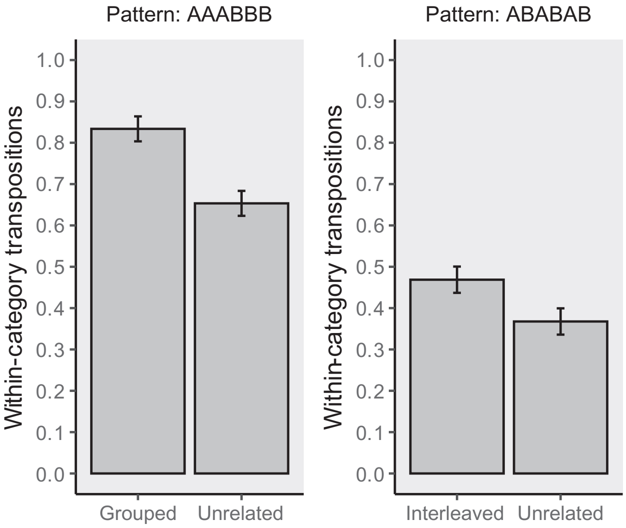
Note. Proportion of correct responses as a function of semantic similarity (grouped, interleaved, dissimilar) and serial position (one through eight). Error bars represent 95% within-subject confidence intervals. See the online article for the color version of this figure.

Experiment 3

Experiment 3 tested the robustness of the impact of semantic similarity on participants' ability to recall the order of items. We manipulated semantic similarity in an order reconstruction task. Contrary to our previous experiments, we switched to a between-subject design. One group of participants was presented with lists composed of either fully similar or fully dissimilar items. Another group of participants was presented with lists composed of either semantically grouped or fully dissimilar items. If the increased order memory performance observed in Experiments 1 and 2 is due to the semantic grouping structure per se, we expect to observe a semantic similarity effect specifically in the group of participants receiving the semantically grouped items. In contrast, no semantic similarity effect on order errors is expected in the group of participants receiving the pure lists, thus replicating previous studies (e.g., Saint-Aubin & Poirier, 1999). We decided to switch to a between-subject design, because the grouping manipulation for one type of list could bias participants toward implementing the same grouping strategy for the nongrouped but semantically similar lists (Bailey et al., 2011; Farrell, 2012; Farrell et al., 2011).

Figure 5

Experiment 2—Increase of Within-Category Transposition Proportions as a Function of List Structure and Semantic Condition



Note. Proportion of within-transposition across semantic conditions. Left panel: Pattern AAABBB, grouped versus dissimilar condition. Right panel: Pattern ABABAB, interleaved versus dissimilar condition. Error bars represent 95% within-subject confidence intervals.

Method

Participants

One hundred twenty adults aged between 18 and 35 participated in this experiment. An optional stopping rule was used for determining optimal sample size (Schönbrodt & Wagenmakers, 2018). We started with a sample size of 30 participants in each group, leading to an initial sample size of 60 participants. In case the BF did not reach a sufficient level of evidence (BF > 10 for either the null or the alternative hypothesis) concerning the critical effects of interest, we planned to recruit 30 more participants in each group. This led to a total sample size of 120 participants (i.e., 60 participants in each experimental group). 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

Stimuli involved 20 categories, each composed of six items. Lists in the grouped condition were built by randomly choosing six items from two different categories, leading to two categories of three items each. The fully similar lists were built by using the items from one category. The dissimilar lists were built by randomly sampling each item from a different semantic category. One group of participants received the fully similar and dissimilar lists. Another group of participants received the grouped and fully dissimilar lists. This way of building the material led to 20 trials per condition. Each item was presented twice across the whole experiment: once in a similar list, and once in a dissimilar list. Items were presented visually so that their presentation matched their representation at test for the reconstruction of order. All other constraints were identical to those described in Experiments 1 and 2.

Procedure

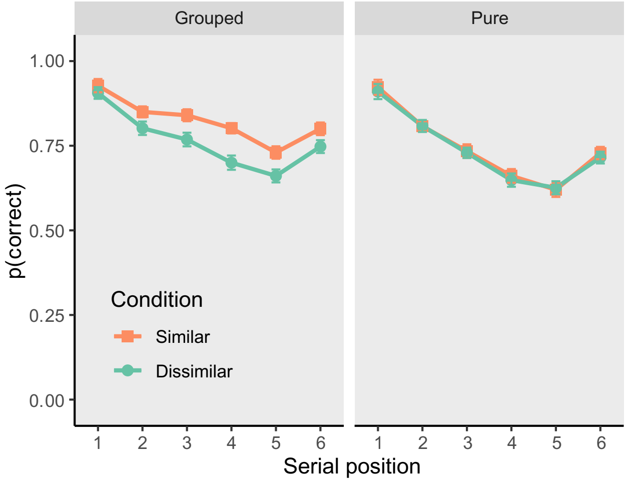
The experimental procedure was identical to Experiment 2, except that set size was now reduced from eight to six to shorten the experiment's duration. Participants performed three training trials before beginning the main experiment. All other aspects of the experiment were identical to Experiment 2, including the scoring procedure and statistical analyses.

Results

The proportion of correct responses as a function of semantic similarity (similar, dissimilar), serial position (one through six), and list structure (grouped, pure) was assessed using a Bayesian repeated-measures ANOVA. The best model was the model including all main effects, the interaction between semantic similarity and list structure, the interaction between serial position and list structure, the random intercept, and the random slopes of semantic similarity and serial position. As compared to the best model, we found anecdotal evidence supporting the main effect of list structure ($BF_{10} = 1.678$) and decisive evidence supporting the main effect of semantic similarity ($BF_{10} = 988.875$). There was very strong evidence supporting the interaction between semantic similarity and list structure ($BF_{10} = 67.894$). Bayesian paired-samples *t* tests indicate that there was decisive evidence supporting the beneficial effect of semantic similarity on order memory in participants receiving the lists of semantically grouped items ($BF_{10} = 1.305e+4$, see Figure 6, left panel). In participants receiving the pure lists, we found

Figure 6

Experiment 3—Serial Position Curves for Recall Performance



Note. Proportion of correct responses as a function of semantic similarity (grouped, dissimilar), list structure (grouped, pure), and serial position (one through six). Error bars represent 95% within-subject confidence intervals. See the online article for the color version of this figure.

moderate evidence against the semantic similarity effect ($BF_{01} = 5.843$, see Figure 6, right panel).

In addition to these results, we also report the pattern of transposition errors, as done in the previous experiments. As this analysis was not the focus of Experiment 3, we decided to report these results in Appendix C. We reproduced the same findings as the previous experiments: semantic similarity constrained transposition errors, but only in the semantically grouped condition.

Discussion

Experiment 3 shows a beneficial impact of semantic similarity on order memory when items are presented in semantic subgroups compared to lists composed of dissimilar items. When semantic similarity was manipulated using pure lists of semantically similar and dissimilar items, no such benefit was observed. This indicates that the beneficial effect of semantic similarity on order memory is specific to cases where semantically similar items form subgroups in the list.

Most likely, this beneficial effect for order recall arises from the category structure imposed by the semantic categories. This structure constrains order memory, preventing transpositions outside of the semantic category more strongly than promoting transpositions within the category, producing a net benefit. This is however difficult to determine only based on the proportion of within-category transpositions, because this score confounds the absolute numbers of within- and between-category transpositions. We, therefore, report in Table 1 the separate proportions for within-category and for between-category transposition errors out of all list items that were recalled, averaged across all participants. There was a massive drop of between-category transpositions in the grouped compared to the dissimilar condition. In contrast, within-category transpositions remained stable. The only exception is Experiment 1, for which there was a slight increase of within-category transpositions (i.e., $0.127 - 0.099 = 0.028$). This slight increase is, however, weaker than the reduction of between-category transpositions (i.e., $0.01 - 0.063 = -0.053$), thus creating a net benefit for order memory, as reported in the results section of Experiment 1.

So far, the preference for within-group transpositions observed across Experiments 1 through 3 were found in conditions in which the list structures were relatively obvious and predictable. Across all experiments, the semantic structures were always of type "AAABBB" or "ABABAB." With such an obvious manipulation, it is possible that participants rapidly developed a long-term knowledge of the list structure. This could have in turn provoked the pattern of within-category transpositions across Experiments 1 through 3. For instance, as soon as participants encode the words "leopard" and "puma" (in this order), they might recognize that the remaining list structure would necessarily follow an "AAABBB" structure. They could use this knowledge to constrain their recall or reconstruction: All words from the A category go into the first three list positions, all words from the B category go into the last three positions. This explanation entails that the constraint of structured semantic similarity on order memory arises from the strategic application of knowledge about the list structure when participants decide about their responses. An alternative is that this constraint arises in a nonstrategic manner from the interaction of semantically similar and dissimilar representations in WM. Experiment 4 provides a test between these two alternatives.

Experiment 4

Experiment 4 tested whether semantic similarity can constrain serial order errors even in conditions in which it is difficult to predict the semantic structure of the list. Semantic similarity was manipulated using lists composed of two semantic categories. Contrary to previous experiments, we included all possible patterns of allocating categories to list positions, thereby making the list structure unpredictable. For instance, given the semantic categories "A" and "B," participants could be presented with a pattern such as "AABBAB," "ABABAB," "ABAABB," "AAABBB," and so forth. A full list of all category patterns is reported in Table 2. Furthermore, we decided to switch back to a serial recall paradigm. We were concerned that with an order reconstruction task, participants would maintain the category pattern in memory, and use it to increase their memory performance, ignoring the items themselves. With a serial recall procedure, such a strategy would be costly as items need to be maintained. If the increased within-category transposition proportion observed across Experiments 1 through 3 is due to participants predicting the category pattern of the list, it should no longer be observed.

Method

Participants

Forty adults aged between 18 and 35 participated in this experiment. Participants were recruited on the online platform Prolific.

Table 1

Patterns of Semantic Structure Used in Experiment 4

	Pattern: AAABBB		Pattern: ABABAB	
Experiment	Grouped	Dissimilar	Interleaved	Dissimilar
Experiment 1				
Within	0.127	0.099	0.059	0.034
Between	0.01	0.063	0.113	0.128
Experiment 2				
Within	0.09	0.092	0.064	0.054
Between	0.022	0.053	0.07	0.091
Experiment 3				
Within	0.071	0.07	N/A	N/A
Between	0.016	0.048	N/A	N/A

Note. Each letter refers to a semantic category.

Table 2

Patterns of Semantic Structure Used in Experiment 4

Patterns of semantic structure					
Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
A	A	A	B	B	B
A	B	A	A	B	B
A	B	A	B	A	B
A	B	A	B	B	A
A	A	B	A	B	B
A	A	B	B	A	B
A	A	B	B	B	A
A	B	B	A	A	B
A	B	B	A	B	A
A	B	B	B	A	A

Note. Each letter refers to a semantic category.

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

Stimuli involved 40 categories, each composed of six items. The similar lists were created by sampling three items from two different semantic categories, resulting in 40 semantically similar lists. The dissimilar lists were created by sampling items from different semantic categories, resulting in 40 semantically dissimilar lists. There were therefore 80 experimental trials across the whole experiment. Each item appeared twice across the whole experiment: once in a similar list, and once in a dissimilar list.

With six items drawn from two distinct semantic categories, it was possible to create 10 different patterns of list structure (Table 2). Each of the 40 semantically similar lists was randomly assigned to one of these 10 patterns, resulting in four lists in each pattern. The position of each item was defined according to the semantic pattern they were assigned to. For instance, a list assigned to the "ABBABA" pattern could be "**lion**, hand, elbow, **cheetah**, leg, **tiger**."

Procedure

Words were visually presented at the center of the screen at a pace of one item/s. Upon retrieval, a box appeared in the middle of the screen, prompting the participants to type their answer. To help participants keep track of their progress, a number below the box indicated the position of the item to be recalled, starting from "1." If participants did not know an item, they could leave the prompt box empty, resulting in an omission error. To submit a response, participants had to press the "Enter" key of their keyboard. This automatically led to the cueing of the next to-be-remembered word. Participants performed three training trials before the beginning of the main experiment.

Scoring Procedure

The scoring procedure was the same as the one used in Experiment 1.

Statistical Analysis

In this experiment, each semantic structure involved only four trials, which made the pattern of transposition errors difficult to analyze when computed as we previously did. This is because the scoring analysis we used so far requires enough trials so that transposition errors would occur at all. If the participant did not produce any transposition error in each condition, the proportion of within-category transposition would be 0/0. To overcome this problem, we analyzed the proportion of transposition errors using a permutation test. For each trial, we computed the number of within and between-category transposition errors according to their semantic structure, resulting in an overall within-category transposition proportion across all trials. This score was compared to a null distribution, which was built by reanalyzing 10^6 times the within-category transposition score after randomly shuffling the semantic labels associated with each trial. Therefore, this resampling process gives the null distribution under which the semantic structure would be randomly assigned for each trial. A *p* value was obtained by computing the proportion of within-category transpositions in the null distribution above the observed proportion of within-category transpositions. Note that the permutation test was performed across all participants aggregated, rather than on each participant. Given that the permutation test requires many computational operations, it was run using the Julia programming language (https://julialang.org/).

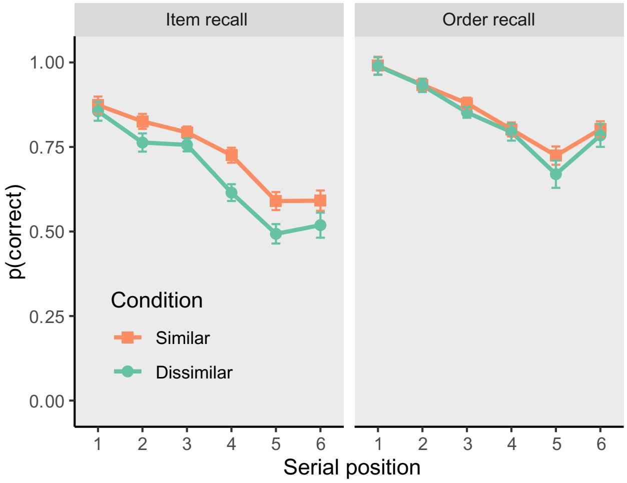
Results

Recall performance as a function of semantic condition (similar, dissimilar) and serial position (one through six) was analyzed using a Bayesian repeated-measures ANOVA. Using an item recall criterion, the best model was the model including all main effects and the interaction term, a random intercept, and the random slope of serial position. Comparing the best model to one removing the fixed main effect of semantic similarity, we found decisive evidence supporting that effect ($BF_{10} = 3.163e+19$). As can be seen in Figure 7, left panel, semantically similar items were better recalled than semantically dissimilar items. The same analysis was conducted using an order recall criterion. The best model was the model including both main effects, a random intercept, and the random slope of serial position. The evidence concerning the main effect of semantic similarity was ambiguous ($BF_{10} = 1.442$). These results are displayed in Figure 7, right panel. Thus, semantic similarity increased item memory, but did not credibly impact order memory. Zooming in on the AAABBB pattern—the only one that could induce grouping—a Bayesian *t* test yields strong evidence for better order memory in the similar than the dissimilar condition ($BF_{10} = 45$).

Next, we tested whether the category structures of the lists affect patterns of transposition errors. Results from the permutation tests are shown in Figure 8. The figure shows, for each similarity

Figure 7

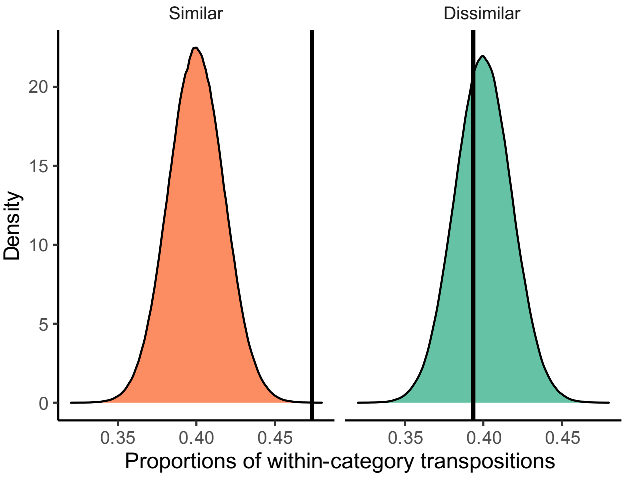
Experiment 4—Serial Position Curves for Recall Performance



Note. Recall performance as a function of semantic condition (similar, dissimilar) and serial position (one through six). Left panel: Item recall criterion. Right panel: Order recall criterion. Error bars represent 95% within-subject confidence intervals. See the online article for the color version of this figure.

Figure 8

Experiment 4—Proportion of Within-Category Transpositions Against a Null Distribution From Permutation Tests



Note. Density of the null distribution of within-category transposition proportions in the permutation tests. The black bar indicates the observed proportion of within-category proportions. See the online article for the color version of this figure.

condition, the distribution of within-category transposition proportions obtained from the permutation test, which reflects the null hypothesis. The black line indicates the observed proportion (i.e., the proportion of within-category transpositions without shuffling the category labels). As can be seen in the left panel, the observed proportion has a very low probability to appear only by chance ($p = 1.5\text{e}{-5}$, $d = 4.147$). For comparison purpose, the same analysis is reported in the semantically dissimilar condition in the right panel. In this analysis, the labels in the observed proportion were randomly assigned, as the dissimilar lists do not have any of the category patterns listed in Table 2. The results from the dissimilar condition show what we should observe if the lists' category pattern did not have any impact on transposition errors: the observed proportion falls in the middle of the distribution ($p = .623$, $d = 0.351$).

It could be argued that the effect observed in Figure 8 is entirely driven by the most obvious list structures, such as the "AAABBB" and "ABABAB" structures. If this is the case, the permutation test should no longer be significant when discarding trials involving these semantic structures. We report in Appendix D a reanalysis showing that the permutation test remained significant under these conservative conditions. Therefore, lists' category pattern constrained transposition errors in the similar condition.

Discussion

Experiment 4 manipulated semantic similarity in such a way that the category structures of lists composed of semantically similar items were unpredictable. We found that semantic similarity improved item memory, while leaving order memory unaffected. Only with the grouped pattern AAABBB, we observed a beneficial effect of semantic similarity on order memory. These results converge with those observed in Experiments 1 through 3, in which a beneficial effect of semantic similarity on order memory appeared only when semantically similar items were arranged in a grouped fashion. Apparently, the beneficial effect of semantic similarity on order memory arises only if the category pattern encourages participants to group the lists.

Within-category transposition analyzed through permutation tests indicated that lists' category structures constrained transposition errors significantly more than what would normally occur by chance. This constraint on transpositions was found regardless of the pattern assigning categories to list positions. These results suggest that the constraint in favor of within-category transposition errors reflects a nonstrategic process and is not driven by people's long-term knowledge of the lists structures.

General Discussion

There were two main findings from this study. First, semantic similarity enhanced order memory, but only when the similar items were presented in a grouped manner, thus replicating previous observations (Kowialiewski et al., 2022). Second, semantic similarity reliably constrained the pattern of order errors, regardless of the lists' semantic structure, confirming the earlier results by Poirier et al. (2015). These results provide robust evidence that semantic similarity can also impact order memory, but in a more specific manner relative to other types of similarity such as phonological similarity.

Potential Objections

One could argue that our semantic manipulations do not involve semantic similarity, but a form of semantic relatedness. It is conceivable that members of the same category are merely related to a superordinate concept but are not similar to each other. Evidence suggests otherwise. First, studies requiring people to generate features from individual concepts have shown that members of the same category share more features than member of different categories (Binder et al., 2016; Devereux et al., 2014), implying that they are indeed similar. Second, concepts drawn from the same semantic category elicit more similar patterns of neural activations than members of different categories (Xu et al., 2018), and this specifically in the anterior temporal lobe, a core neural region involved in semantic processing (Ralph et al., 2017). Third, in delayed recall paradigms, release from proactive interference typically occurs when switching materials from one semantic category to another between trials, implying that members of the same category interfere with each other through their shared features (Craik & Birtwistle, 1971; Wickens, 1970). Taken together, members of a category share more semantic features than members of different categories, they have more similar neural representations, and they are confused more with each other in episodic memory. We are therefore confident that category membership constitutes a valid manipulation of semantic similarity. Finally, Neath et al. (2022) showed that order errors are consistently unaffected by semantic similarity manipulated in several ways, including synonyms, an extreme form of semantic similarity. These results provide convergent evidence for our conjecture that semantic similarity does not negatively affect order memory in immediate tests of memory for order.

Are Semantic Features Encoded in WM?

Most models of WM maintain order information via a binding mechanism associating item features to context (Burgess & Hitch, 1999, 2006; Lewandowsky & Farrell, 2008; Oberauer et al., 2012).

When retrieving an item from its context, this leads to the retrieval of a degraded representation of the original item. To produce a legitimate response, this degraded representation must be compared to a set of retrieval candidates, a process called *redintegration* (Schweickert, 1993). As the similarity between the retrieval candidates and the target item increases (i.e., in a list composed of similar items), the probability to select another item than the target one also increases, resulting in more confusion errors. This set of assumptions explains why phonologically similar items are confused more often with each other than phonologically dissimilar items (Baddeley, 1966; Fallon et al., 2005; Gupta et al., 2005; L. Nimmo & Roodenrys, 2005; Roodenrys, Guitard, et al., 2022). The fact that semantic similarity does not increase confusion errors (Kowialiewski, Krasnoff, et al., 2023; Neath et al., 2022; Poirier & Saint-Aubin, 1995; Saint-Aubin & Poirier, 1999) suggests that semantic features are not included in the redintegration process. In contrast to this assumption, the present study provides strong evidence that semantic similarity constrains transposition errors.

We are therefore faced with an apparent contradiction. On the one side, there is robust evidence showing that when semantic similarity is manipulated at the whole-list level, semantic features are not used for the selection of retrieval candidates during recall. On the other side, when semantic similarity is manipulated using items drawn from different semantic categories, we do find evidence suggesting that semantic features are used to constrain which retrieval candidates are chosen as responses at each list position. The findings reported in Table 1 with lists composed of items from two categories show that transpositions are unlikely to cross-category boundaries, implying that information about which category was in each list position is used during recall. Why, then, is that same information not used to reduce the number of transposition errors overall in a purely dissimilar list, relative to a purely similar list? These results challenge all current models of serial order memory for verbal items. In the next section, we evaluate the plausibility of several mechanisms that could explain semantic similarity effects.

Possible Mechanisms

The Interactive Activation Model

Recently, it has been shown that semantic similarity effects in serial recall can be simulated by an architecture integrating spreading of activation principles (Collins & Loftus, 1975; Dell et al., 1997) in a semantic network (Kowialiewski & Majerus, 2020; Kowialiewski, Lemaire, & Portrat, 2021). In this model, semantic features are not bound to context. Instead, the semantic similarity advantage is explained by semantically similar items reactivating each other in the semantic network through interactive activation, resulting in stronger sustained activation. In many models of WM, if an item's activation is below a retrieval threshold, the item is omitted. Thanks to their higher activation level, semantically similar items overcome this omission threshold more frequently than semantically dissimilar items. Simulations have shown that this model predicts a recall advantage for semantically similar versus dissimilar items. Because in this model the semantic features of items are not bound to context, this leads to zero impact on order memory, which is an accurate prediction as long as semantic similarity is manipulated using pure lists of similar or dissimilar items. Another logical consequence of not binding semantic features to context is that this model cannot predict the semantic constraint on transposition errors observed in the present study. Therefore, in its current form, the interactive activation model only partially captures the semantic similarity effect. A more extreme version of this account suggests that item order can be maintained via the pattern of activation in a semantic network (i.e., the ANet account, see Poirier et al., 2015). Recent simulations have shown the implausibility of such a model to account for semantic similarity effects (Kowialiewski, Lemaire, Majerus, & Portrat, 2021), including the data reported by Poirier et al. (2015).

The Feature Model

In the Feature Model (Nairne, 1990), as well as its revised version (i.e., the Revised Feature Model or RFM, see Poirier et al., 2019; Saint-Aubin et al., 2021), items are represented in a distributed fashion using features coding each item's dimension (e.g., phonology, orthography, …). Items are temporarily stored in WM using a copy of the original ones. One important source of forgetting is *feature overwriting*, a process whereby features of the currently encoded item N overwrite (i.e., setting their values to zero) the overlapping features of the preceding item $N - 1$ (or all preceding items $N - i$, in the RFM). Due to overwriting, items become weaker when followed by other similar compared to dissimilar items. At retrieval, the partially overwritten traces stored in WM are redintegrated by comparing them to those in long-term memory.[5] If semantic features are represented in WM, the model can explain the influence of the semantic structure of a memory list on transposition errors: During redintegration, a trace in WM is less likely to be confused with a semantically dissimilar than with a semantically similar representation in long-term memory. At the same time, the model also predicts *decreased* order memory for semantically similar compared to dissimilar lists, an unobserved result. In addition, due to the feature overwriting mechanism, this model predicts decreased item memory for semantically similar lists, which is again in contradiction with the empirical data. The semantic similarity effect poses therefore a challenge for the feature model, along with the recent evidence arguing against feature overwriting (Roodenrys, Miller, & Josifovski, 2022). If the assumptions of the RFM are revised by not including the feature overwriting mechanism, then it could potentially account for the semantic similarity effect using the category encoding assumption described below.

A Grouping Mechanism

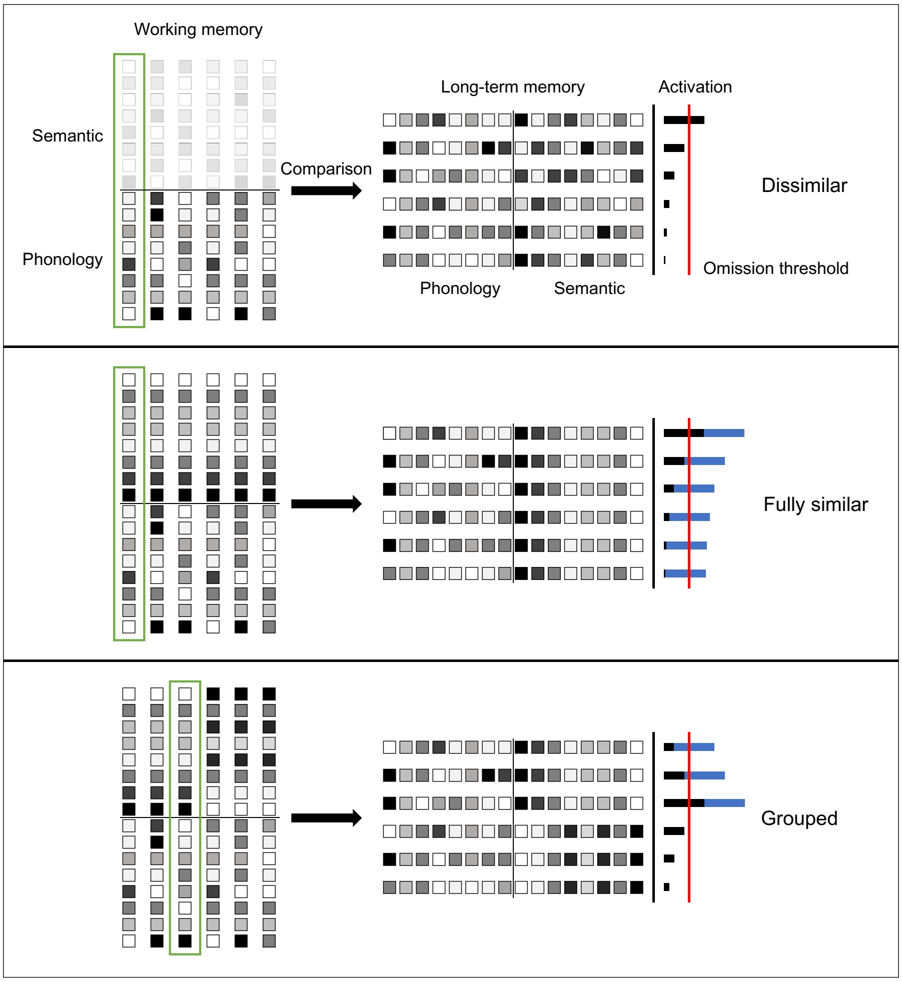
Another potential mechanism relates to grouping. Temporal grouping is a well-established effect in the serial recall literature (Burgess & Hitch, 1999, 2006; Farrell, 2012; Ryan, 1969a, 1969b). Grouping is induced by inserting a short pause between two successive items (e.g., ABC—pause—DEF). As compared to nongrouped lists, temporal grouping is characterized by better recall performance, both at the item and order level (Hitch et al., 1996; Kowialiewski, Gorin, & Majerus, 2021; Ng & Maybery, 2002; Parmentier et al., 2006). Temporal grouping increases within-group transpositions and reduces between-group transpositions. The temporal grouping effect is modeled by implementing a context

[5] In the (Revised) Feature Model, WM is referred to as "primary memory," and long-term memory is referred to as "secondary memory." As these are just less common terms for WM and long-term memory, respectively, we use the more common terms here.

representing serial position on two different levels: one at the group level, and one at the within-group level. When a temporal pause occurs, a new group marker is used, and the position of subsequent to-be-encoded items is coded relatively to this new group (i.e., the first item in the second group is bound to the second group, and the first within-group position). At retrieval, items are cued using a position context consisting of both the group markers and the within-group markers. In a temporally grouped context, the grouping markers mechanically restrict the between-item competition to those items associated with the currently used grouping context, thus

Figure 9

Illustration of the Category-Encoding Assumption



Note. In this model, items are encoded into WM in a distributed fashion using feature vectors. Each vector contains both phonological and semantic representations of items. At retrieval, one vector is selected in WM, and this vector is then compared to a set of retrieval candidates. In immediate serial recall, the retrieval candidates are the items stored in long-term memory. This comparison process leads to a degree of activation associated with each retrieval candidate. Highly activated items are more likely to be selected. If an item's activation is below the omission threshold (illustrated in red), it will not be recalled and the model produces an omission error. Upper panel: In a semantically dissimilar list, items' semantic features are not kept in WM (indicated by their transparency), leading to no contribution of semantics during the comparison stage. Middle panel: In lists of pure semantically similar lists, only the items' shared features are kept in WM. This can be seen by the redundant semantic feature vectors, which are shared by all memoranda. During the comparison stage, these additional features will add a constant boost of activation to all items. This additional boost allows items to overcome the omission threshold more often, thus leading to better item memory. Since the category adds a constant boost of activation to all memoranda, their relative activation level remains the same as compared to a purely dissimilar condition, leading to no impact on order memory. Lower panel: When items are presented in subgroups composed of two categories, the additional semantic features will benefit only those items sharing the same semantic features (i.e., those from the same semantic category). This automatically restricts the set of candidates to those sharing the same semantic category, thus preventing cross-category transposition errors. WM = working memory. See the online article for the color version of this figure.

increasing order memory, and reducing between-group transposition errors. The temporal grouping effect is strikingly comparable to the results observed in the present study. It is therefore appealing to call for an equivalent explanation for both phenomena.

A grouping explanation is plausible for the category pattern AAABBB, which could motivate participants to impose a group structure on their list representation. It is less plausible, however, for the interleaved pattern ABABAB, as the list cannot be broken down into two groups along category lines. Still, one could assume, in analogy to models of grouping, a two-level positional context in which the positions with items from the same category—here, the odd versus the even positions—receive a common representation akin to a group marker. Even this extended grouping explanation, however, is unlikely for the unpredictable category patterns that we used in Experiment 4: To use such a two-layer position context as effective retrieval cue, people would have to know the pattern of a list. For instance, when a list uses the pattern AABABB, positional contexts 1, 2, and 4 share one group-like marker, and positions 3, 5, and 6 share the other group-like marker. Unless the person knows this, they are unable to reproduce the contexts to which each list item is bound at the time of test, and therefore could not use the group-like part of the position context to constrain retrieval. Therefore, this grouping idea poses more conceptual problems than it solves.

Category Encoding

A mechanism we are currently exploring through computational modeling is one in which WM retains categorical information. We illustrate the way this model works in Figure 9. At list presentation, WM encodes both phonological and semantic features. When people are presented with semantically similar items (e.g., "leopard, cheetah, lion"), only the features shared by all these items survive (i.e., has fur, is dangerous, is a big cat, is a wild animal). This mechanism could be formally implemented by assuming that items' shared features reactivate and support each other via interactive activation, leaving unshared features to die down. This mechanism has the consequence of adding a constant representation to all semantically similar items (i.e., all items share the same semantic features). When items are semantically dissimilar, no features survive, because dissimilar items share little or no features. Items in a semantically dissimilar list would therefore only be stored at the phonological level—their semantic features will all be set to a neutral value (e.g., zero). Due to the additional semantic features, WM representations for semantically similar items will be richer and redintegration will be facilitated, leading to a recall advantage for semantically similar items.

With this representational assumption, lists of purely similar items and lists of purely dissimilar items will both have item representations with all identical semantic features—the shared feature values in the former case, and the neutral (i.e., zero) feature values in the latter. Therefore, in both purely similar and purely dissimilar lists, semantic features are equally useless for discriminating between list items—and hence, for reducing order errors. However, in lists composed of items from two categories, items will have semantic features shared with other items of the same category but that allows to discriminate them from the items of the other category. For instance, in a list of pattern AABABB, the items from category A will have one set of semantic features (those shared by all members of the A category), and the items from category B will have another set of semantic features (those shared by members of the B category). This in turn will help prevent cross-category transpositions. The reason why this reduction of cross-category transposition only enhances order memory for the AAABBB pattern remains to be understood. One possibility is that this pattern is the one where the detection of semantic relationships is the most obvious, as also shown by the increased item memory for the grouped versus interleaved pattern reported in Experiment 1 (i.e., the separation effect). This may in turn reduce more strongly cross-category transpositions than in any other list structure, leading to a net benefit at the order level. We reach a point where the use of computational modeling is necessary, otherwise, these interpretations would remain vague and imprecise.

Conclusion

Semantic similarity poses a challenge to theories of memory for serial order. The effects of semantic similarity are qualitatively different from those of other similarity dimensions—in particular phonological similarity. Findings from comparisons of purely similar and purely dissimilar lists seem to imply that semantic information plays no role in immediate memory for order, whereas findings from lists composed of two semantic categories—as in the present experiments—seem to imply the opposite. Currently, no model of serial order memory can account for the full set of known effects of semantic similarity. However, a computational instantiation of our proposed category encoding mechanism holds promise for developing a coherent and complete account of the apparently contradictory effects we have investigated in this article and that other memory scientists have reported elsewhere.

References

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

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158–175. https://doi.org/10.3758/s13428-011-0123-7

Bailey, H., Dunlosky, J., & Kane, M. J. (2011). Contribution of strategy use to performance on complex and simple span tasks. *Memory & Cognition*, *39*(3), 447–461. https://doi.org/10.3758/s13421-010-0034-3

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3–4), 130–174. https://doi.org/10.1080/02643294.2016.1147426

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. https://doi.org/10.1037/0033-295X.106.3.551

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

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press.

Craik, F. I., & Birtwistle, J. (1971). Proactive inhibition in free recall. *Journal of Experimental Psychology*, *91*(1), 120–123. https://doi.org/10.1037/h0031835

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

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127. https://doi.org/10.3758/s13428-013-0420-4

Fallon, A. B., Mak, E., Tehan, G., & Daly, C. (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. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271. https://doi.org/10.1037/a0027371

Farrell, S., Wise, V., & Lelièvre, A. (2011). Relations between timing, position, and grouping in short-term memory. *Memory & Cognition*, *39*(4), 573–587. https://doi.org/10.3758/s13421-010-0053-0

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), Article e0168699. https://doi.org/10.1371/journal.pone.0168699

Guérard, K., Tremblay, S., & Saint-Aubin, J. (2009). Short Article: Similarity and binding in memory: Bound to be detrimental. *Quarterly Journal of Experimental Psychology*, *62*(1), 26–32. https://doi.org/10.1080/17470210802215277

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

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. https://doi.org/10.1006/cogp.1998.0685

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

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 116–139. https://doi.org/10.1080/713755609

Ishiguro, S., & Saito, S. (2020). The detrimental effect of semantic similarity in short-term memory tasks: A meta-regression approach. *Psychonomic Bulletin & Review*, *28*(2), 384–408. 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

Karlsen, P. J., Imenes, A. G., Johannessen, K., Endestad, T., & Lian, A. (2007). Why does the phonological similarity effect reverse with nonwords? *Psychological Research*, *71*(4), 448–457. https://doi.org/10.1007/s00426-005-0042-2

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*, *47*(12), 1958–1970. https://doi.org/10.1037/xlm0001031

Kowialiewski, B., Krasnoff, J., Mizrak, E., & Oberauer, K. (2022). The semantic relatedness effect in serial recall: Deconfounding encoding and recall order. *Journal of Memory and Language*, *127*, Article 104377. https://doi.org/10.1016/j.jml.2022.104377

Kowialiewski, B., Krasnoff, J., Mizrak, E., & Oberauer, K. (2023). Verbal working memory encodes phonological and semantic information differently. *Cognition*, *233*, Article 105364. https://doi.org/10.1016/j.cognition.2022.105364

Kowialiewski, B., Lemaire, B., Majerus, S., & Portrat, S. (2021). Can activated long-term memory maintain serial order information? *Psychonomic Bulletin & Review*, *28*(4), 1301–1312. https://doi.org/10.3758/s13423-021-01902-3

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*, Article 104208. https://doi.org/10.1016/j.jml.2020.104208

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

Kowialiewski, B., Majerus, S., & Oberauer, K. (2023). *Semantic order categories* [Data set]. https://osf.io/wzndt

Lewandowsky, S. (1999). Redintegration and response suppression in serial recall: A dynamic network model. *International Journal of Psychology*, *34*(5–6), 434–446. https://doi.org/10.1080/002075999399792

Lewandowsky, S., & Farrell, S. (2008). Short-term memory: New data and a model. *Psychology of Learning and Motivation*, *49*, 1–48. https://doi.org/10.1016/S0079-7421(08)00001-7

Lian, A., & Karlsen, P. J. (2004). Advantages and disadvantages of phonological similarity in serial recall and serial recognition of nonwords. *Memory & Cognition*, *32*(2), 223–234. https://doi.org/10.3758/BF03196854

Lin, P.-H., & Luck, S. J. (2009). The influence of similarity on visual working memory representations. *Visual Cognition*, *17*(3), 356–372. https://doi.org/10.1080/13506280701766313

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*, Article 357. https://doi.org/10.3389/fnhum.2013.00357

Majerus, S. (2019). Verbal working memory and the phonological buffer: The question of serial order. *Cortex*, *112*, 122–133. https://doi.org/10.1016/j.cortex.2018.04.016

Majerus, S., Attout, L., Artielle, M.-A., & Van der Kaa, M.-A. (2015). The heterogeneity of verbal short-term memory impairment in aphasia. *Neuropsychologia*, *77*, 165–176. https://doi.org/10.1016/j.neuropsychologia.2015.08.010

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

Martin, N., & Saffran, E. M. (1997). Language and auditory-verbal short-term memory impairments: Evidence for common underlying processes. *Cognitive Neuropsychology*, *14*(5), 641–682. https://doi.org/10.1080/026432997381402

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

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

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), 196–209. https://doi.org/10.1027/1618-3169/a000557

Ng, H. L. H., & Maybery, M. T. (2002). Grouping in short-term verbal memory: Is position coded temporally? *The Quarterly Journal of Experimental Psychology Section A*, *55*(2), 391–424. https://doi.org/10.1080/02724980143000343

Nimmo, L., & Roodenrys, S. (2005). The phonological similarity effect in serial recognition. *Memory*, *13*(7), 773–784. https://doi.org/10.1080/09658210444000386

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

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin*, *143*(9), 992–1009. https://doi.org/10.1037/bul0000108

Norris, D. (2019). Even an activated long-term memory system still needs a separate short-term store: A reply to cowan (2019). *Psychological Bulletin*, *145*(8), 848–853. https://doi.org/10.1037/bul0000204

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. *Psychology of Learning and Motivation*, *51*, 45–100. https://doi.org/10.1016/S0079-7421(09)51002-X

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

Papagno, C., Comi, A., Riva, M., Bizzi, A., Vernice, M., Casarotti, A., Fava, E., & Bello, L. (2017). Mapping the brain network of the phonological loop: The Phonological Loop Brain Network. *Human Brain Mapping*, *38*(6), 3011–3024. https://doi.org/10.1002/hbm.23569

Parmentier, F. B. R., Andrés, P., Elford, G., & Jones, D. M. (2006). Organization of visuo-spatial serial memory: Interaction of temporal order with spatial and temporal grouping. *Psychological Research Psychologische Forschung*, *70*(3), 200–217. https://doi.org/10.1007/s00426-004-0212-7

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., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *50*(4), 408–412. https://doi.org/10.1037/1196-1961.50.4.408

Poirier, M., Saint-Aubin, J., Mair, A., Tehan, G., & Tolan, A. (2015). Order recall in verbal short-term memory: The role of semantic networks. *Memory & Cognition*, *43*(3), 489–499. https://doi.org/10.3758/s13421-014-0470-6

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

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Roodenrys, S., Guitard, D., Miller, L. M., Saint-Aubin, J., & Barron, J. M. (2022). Phonological similarity in the serial recall task hinders item recall, not just order. *British Journal of Psychology*, *113*(4), 1100–1120. https://doi.org/10.1111/bjop.12575

Roodenrys, S., Miller, L. M., & Josifovski, N. (2022). Phonemic interference in short-term memory contributes to forgetting but is not due to overwriting. *Journal of Memory and Language*, *122*, Article 104301. https://doi.org/10.1016/j.jml.2021.104301

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Ryan, J. (1969a). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology*, *21*(2), 137–147. https://doi.org/10.1080/14640746908400206

Ryan, J. (1969b). Temporal grouping, rehearsal and short-term memory. *Quarterly Journal of Experimental Psychology*, *21*(2), 148–155. https://doi.org/10.1080/14640746908400207

Saint-Aubin, J., Guérard, K., Chamberland, C., & Malenfant, A. (2014). Delineating the contribution of long-term associations to immediate recall. *Memory*, *22*(4), 360–373. https://doi.org/10.1080/09658211.2013.794242

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*, 171–177. https://doi.org/10.3758/BF03196364.

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. 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*, Article 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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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. (2009). The role of associative strength in the semantic relatedness effect on immediate serial recall. *Memory*, *17*(8), 874–891. https://doi.org/10.1080/09658210903376250

Tse, C.-S. (2010). A negative semantic similarity effect on short-term order memory: Evidence from recency judgements. *Memory*, *18*(6), 638–656. https://doi.org/10.1080/09658211.2010.499875

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

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), Article e56. https://doi.org/10.1371/journal.pbio.0050056

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, *77*(1), 1–15. https://doi.org/10.1037/h0028569

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

Xu, Y., Wang, X., Wang, X., Men, W., Gao, J.-H., & Bi, Y. (2018). Doctor, teacher, and stethoscope: Neural representation of different types of semantic relations. *The Journal of Neuroscience*, *38*(13), 3303–3317. https://doi.org/10.1523/JNEUROSCI.2562-17.2018

(*Appendices follow*)

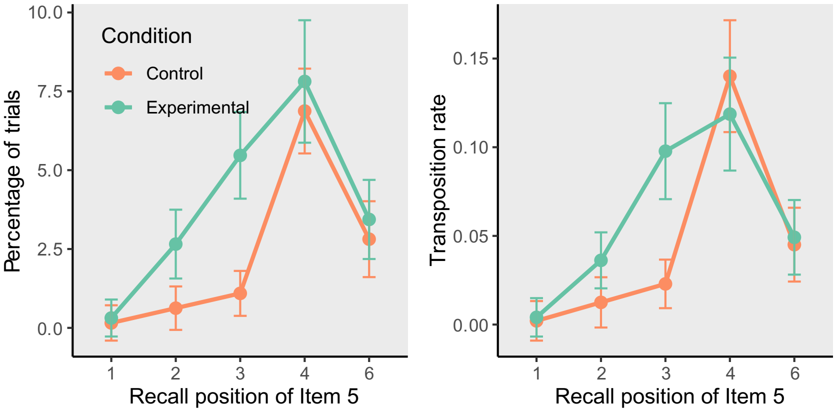
Appendix A

In this section, we report a reanalysis of Experiment 1 from Poirier et al. (2015). In this experiment, participants were presented with six-word lists in which the three first items were semantically similar. In a control condition, the three last items were semantically dissimilar (e.g., officer, badge, siren, yellow, music, tourist). In an experimental condition, the fifth item was semantically similar to the three first items (e.g., officer, badge, siren, fence, police, tractor). The left panel of Figure A1 shows the data as reported in the original study. Poirier et al. used as dependent variable the number of times item five migrated toward the position of another item, proportionalized by the total number of trials (expressed in percentage). As can be seen, results show that item five tended to be transposed more often toward positions 2 and 3.

We reanalyzed these data, by using instead the number of times item five migrated, out of the total number of transposition errors. Results of this reanalysis are shown in the right panel of Figure A1. We ran a Bayesian repeated-measures ANOVA with semantic condition (control, experimental) and recall position (1, 2, 3, 4, 6). The best model was the model including the effect of position, the interaction term, and the random slope of position. As compared to the best model, there was strong evidence supporting the interaction ($BF_{10} = 19.12$). Therefore, the results reported by Poirier and colleagues do not depend on the choice of dependent variable.

Figure A1

Reanalysis of the Results From Poirier et al. (2015)



Note. Left panel: Number of times item number five migrated toward another position, out of the number of trials, as initially reported by Poirier and colleagues. Right panel: Number of times item number five migrated toward another position, out of the total number of transposition errors. Error bars represent 95% within-subject confidence intervals. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix B

Table B1
Detailed Bayes Factors Across All Experiments

Experiment	Criterion	Effect	BF_{10}
Experiment 1	Item recall	Semantic similarity	2.153e+21
		Serial position	2.837e+30
		Interaction	8.646e+25
	Order recall	Semantic similarity	4.031
		Serial position	1.0971e+28
		Interaction	61.989
	Within-category transpositions	AAABBB	2.491e+8
		ABABAB	149
Experiment 2	Order recall	Semantic similarity	4.653e+4
		Serial position	3.442e+35
		Interaction	56
	Within-category transpositions	AAABBB	1.840e+4
		ABABAB	13
Experiment 3	Order recall	List structure	1.678
		Semantic similarity	988.875
		Serial position	3.564e+94
		Semantic Similarity × Serial Position	1/16.086
		Semantic Similarity × List Structure	67.894
		List Structure × Serial Position	431.094
		Triple interaction	1/4.016
	Within-category transpositions	Semantic similarity	3.173
		List structure	329.847
		Interaction	1.049e+4
Experiment 4	Item recall	Semantic similarity	3.163e+19
		Serial position	4.762e+22
		Interaction	227.201
	Order recall	Semantic similarity	1.442
		Serial position	3.909e+28
		Interaction	1/16.697

Note. BF = Bayes factor.

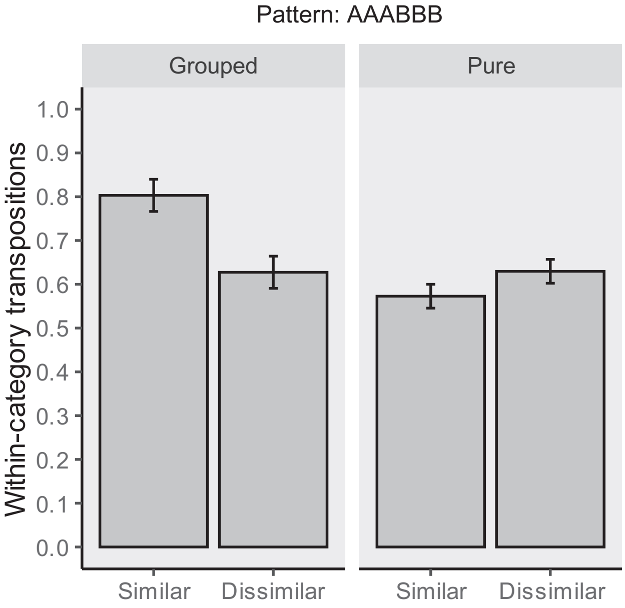
(*Appendices continue*)

Appendix C

We analyzed within-category transposition proportions as a function of semantic similarity (similar, dissimilar) and list structure (grouped, pure) using a Bayesian repeated-measures ANOVA. The best model was the model including all main effects and the interaction. As compared to the best model, we found moderate evidence supporting the semantic similarity effect ($\mathrm{BF}_{10} = 3.173$). We found decisive evidence supporting the effect of list structure ($\mathrm{BF}_{10} = 329.847$) and the interaction term ($\mathrm{BF}_{10} = 1.049\mathrm{e}{+}4$). Bayesian paired-sample *t* tests indicate that whereas no credible evidence supported the effect of semantic similarity for pure lists ($\mathrm{BF}_{10} = 1.05$), there was decisive evidence supporting an effect of semantic similarity in the group of participants receiving the grouped lists ($\mathrm{BF}_{10} = 1{,}558.321$). These results are displayed in Figure C1.

Figure C1

Experiment 3—Increase of Within-Category Transposition Proportions as a Function of Semantic Condition and List Structure



Note. Proportion of within-transposition across semantic conditions. Left panel: Pattern AAABBB, grouped versus dissimilar condition. Right panel: Pattern ABABAB, interleaved versus dissimilar condition. Error bars represent 95% within-subject confidence intervals.

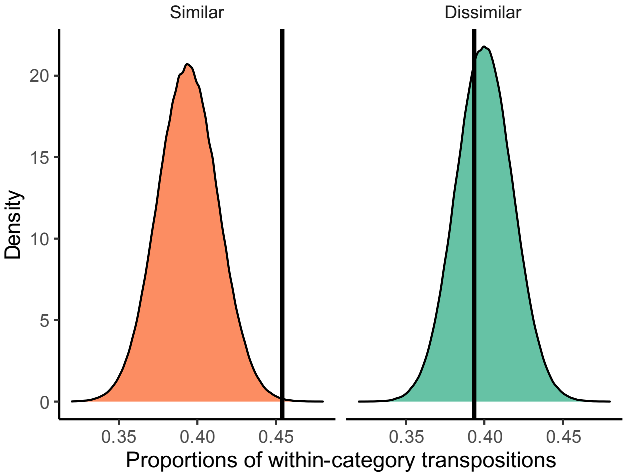
(*Appendices continue*)

Appendix D

We reran the permutation tests from Experiment 4, discarding the trials involving the "AAABBB" and "ABABAB" lists, which are the most obvious and predictable semantic structures. As can be seen in Figure D1, the proportion of within-category transpositions in the unshuffled ratio has an extreme value compared to what would be expected only by chance ($p = 8.53\text{e}{-5}$, $d = 3.12$). In the dissimilar condition, this was not observed ($p = .623$, $d = 0.352$).

Figure D1

Reanalysis of Experiment 4—Proportion of Within-Category Transpositions Against a Null Distribution From Permutation Tests



Note. Density of the within-category proportions null distribution in the permutation tests. The black bar indicates the observed proportion of within-category proportions. See the online article for the color version of this figure.

Received January 25, 2023
Revision received June 21, 2023
Accepted June 22, 2023 ■