Serial position-dependent false memory effects

Halle R. Dimsdale-Zucker, Kristin E. Flegal, Alexandra S. Atkins & Patricia A. Reuter-Lorenz

To cite this article: Halle R. Dimsdale-Zucker, Kristin E. Flegal, Alexandra S. Atkins & Patricia A. Reuter-Lorenz (2018): Serial position-dependent false memory effects, Memory, DOI: 10.1080/09658211.2018.1513039

To link to this article: https://doi.org/10.1080/09658211.2018.1513039

Published online: 28 Aug 2018.

Article views: 38

View Crossmark data
ABSTRACT
Evidence for false recognition within seconds of encoding suggests that semantic-associative influences are not restricted to long-term memory, consistent with unitary memory accounts but contrary to dual store models. The present study sought further relevant evidence using a modified free recall converging associates task where participants studied 12-item lists composed of 3 semantically distinct quartets (sublists) related to a separate, non-presented theme word (i.e., words 1–4/theme1, 5–8/theme2, and 9–12/theme3). This list construction permits assessment of false recall errors from each sublist, and, particularly, the primacy and recency sublists that have been linked to long- and short-term memory stores. Experiment 1 tested immediate free recall for items. Associative false memories were evident from all sublists, however, significantly less so from the recent sublist, which also showed the highest levels of veridical memory. By inserting a brief (3 s) distractor prior to recall, Experiment 2 selectively reduced veridical memory and increased false memory for the recent sublist while leaving the primacy sublist unaffected. These recall results converge with prior evidence indicating the immediacy of false recognition, and can be understood within a unitary framework where the differential availability of verbatim features and gist-based cues affect memory for primacy and recency sublists.

Our memories can mislead us in many ways (for a comprehensive review, see (Gallo, 2006) revealing the fundamentally constructive nature of memory (Schacter & Addis, 2007). Laboratory investigations of memory failures or false memories (FM) date back as early as Bartlett’s seminal War of the Ghosts study (1932), although modern interest in the topic is linked to the popularity of the converging associates or Deese-Roediger-McDermott (DRM) task (Deese, 1959; Roediger & McDermott, 1995). In this task, participants study lists of 12–15 semantic associates that converge on a semantically related but non-presented theme word (e.g., sour, candy, sugar, bitter, good, taste, tooth, nice, honey, soda, chocolate, heart, cake, tart, pie, related to the theme SWEET; for a similar paradigm in phonologically related words, see (Sommers & Lewis, 1999). In this widely replicated effect, participants are highly likely to falsely recall the theme word and can be as confident in these semantic errors as they are in their correct responses.

Prominent explanations of this finding are typically cast within a long-term memory framework and hinge on the idea that FM arise because they fit with the general gist or schema of the studied memoranda (Bransford & Franks, 1971). Support for this view comes from evidence that the degree of interrelatedness among themes and their associates, often referred to as a theme’s backward associative strength (BAS; (Arndt, 2012; Roediger, Watson, McDermott, & Gallo, 2001), influences the likelihood that an associate will be freely generated given a theme word. Additional support comes from the finding that spreading associative activation among items can activate related concepts, semantic schemas (Roediger et al., 2001) or gist (Brainerd & Reyna, 2002), even if these associative processes are implicit (Underwood, 1965).

Recent work has extended the false memory phenomenon to show that memory distortions can also occur with 3- or 4-item lists tested within seconds of encoding (Atkins & Reuter-Lorenz, 2008; Coane, McBride, Raulerson, & Jordan, 2007; Festini & Reuter-Lorenz, 2013). Whether these short lists are tested immediately or after a 20-minute delay, corrected false recognition rates and phenomenological measures of false memory quality (confidence, remember/know) can be approximately equivalent (Flegal, Atkins, & Reuter-Lorenz, 2010). Demonstrations of false memories in canonical short-term or working memory tasks challenge traditional notions that semantic distortions are phenomena unique to long-term memory. They are also consistent with memory models that posit unitary stores and common processes that operate across delays.
More generally, unitary models of memory favour the idea that differences in memory performance across delays can arise from differences in the features or codes of the memory trace that are accessible at the time of retrieval (Jonides et al., 2008; Nairne, 2002; Ranganath & Blumenfeld, 2005). Accordingly, the dynamic interplay between veridical and FM entails the changing strengths of verbatim, perceptual codes and semantic, gist codes (Brainerd & Reyna, 2002). Fuzzy Trace Theory (FTT) posits that verbatim codes can be used to monitor and suppress gist codes (Brainerd & Reyna, 2002), and, as Flegal and Reuter-Lorenz (2014) have discussed, their differential availability across delays could influence delay-dependent differences in the incidence of false memory errors. With these ideas in mind, here we revisit the classic list-learning paradigm with two goals. The first is to ascertain whether false recall errors occur in association with recent sublist; that is, can we extend the evidence for false “short-term memories” to free recall? The second goal is to measure the relative incidence of veridical versus false memories arising from early and late serial positions to assess whether an account such as FTT can provide a unitary and parsimonious explanation for any position-dependent effects that may be observed.

Traditionally the list-learning task was used to dissociate putative short-term and long-term memory stores (reviewed in Capitani, Della Sala, Logie, & Spinnler, 1992). The primacy effect, the benefit for early list items, is relatively resilient to interference (Glanzer & Cunitz, 1966), benefits from slower presentation rates (Murdock, 1962), and is predominately susceptible to semantic errors (Craik & Levy, 1970). The recency effect, the benefit for late list items, can be abolished in the face of interference (Glanzer & Cunitz, 1966), is unchanged based on list presentation rate (Murdock, 1962), and is selectively susceptible to acoustic or phonological errors (Craik, 1968). Thus, the traditional interpretation relevant here is that recall from early and late positions depends on separate long-term and short-term memory stores, respectively. In contrast, unitary models of memory would posit that these memory differences arise from differences in the features of memory traces available for list items at retrieval (Nairne, 2002; Surprenant & Neath, 2009).

Previous work with the list-learning task identified dissociable primacy and recency effects by examining errors associated with words in initial and late positions in lists of otherwise unrelated words (Deese & Kaufman, 1957; Glanzer & Cunitz, 1966; Murdock, 1962). In the present investigation, we created 12-item lists that were composed of three semantically distinct quartets (hereafter called sublists; see Appendix). This design allowed us to track veridical and false memories associated with early (“A”), middle (“B”), and late (“C”) sublists that are related both to the semantic-associative activation among adjacent members in a quartet as well as the established primacy and recency effects from the list considered as a whole.

The present approach builds on previous studies of delay-invariant false memory effects in several ways. In the hybrid design developed by Flegal et al. (2010), participants study associated 4-word lists. For a random subset, recognition is tested immediately after list presentation whereas for the remaining subset participants make an arbitrary motoric response but then receive a surprise recognition test on these items at the end of the experiment. Therefore, with the Flegal paradigm, performance on the long-term test reflects incidental learning in that, unless instructed otherwise (e.g., Flegal & Reuter-Lorenz, 2014), participants likely engage encoding strategies intended simply to support memory for the few seconds preceding the immediate recognition test. In the current paradigm, participants encode items knowing they will be required to recall the entire list. This requirement could promote different encoding strategies for items presented early versus late in the list. Furthermore, the present task uses recall which allows for a more complete and nuanced assessment of memory than recognition testing (Kahana, Howard, & Polyn, 2008). In particular, the relative incidence of veridical versus false recall can provide information about the nature of the cues or features from memory traces arising from different portions of each list that are available at retrieval.

We predicted that semantic FM would be present from the recency sublist but reduced relative to those from the primacy sublist based on the differential availability of gist versus verbatim cues depending on list position (Flegal et al., 2010). Additionally, in Experiment 2 we predicted that inserting a distracting task immediately after the list would reduce the availability of verbatim cues and selectively affect performance from the recency sublist, increasing rates of FM and decreasing veridical responses, while having minimal impact on A or B sublists.

**Experiment 1**

**Method**

**Participants**

Fifty-two University of Michigan undergraduate students took part in the experiment for either course credit or monetary compensation ($10 per hour). The sample size was set *a priori* to ensure that each level of the counterbalance order had the same number of participants. All participants were native English speakers, right-handed, and free from any reported neurological or psychological conditions. Four participants were excluded for reading the words aloud at study, three for experimenter error and equipment malfunctions, two for responding strategies that significantly changed the time demands of the task (perseverative repetition of list items [more than four repetitions per trial], a severe speech impediment), and one for being on psychoactive medications. The following analyses represent data from the remaining 42 participants (14 females; *M* age = 19.45 years, *SD* = 1.11).
**Materials**

One 12-item list was presented on each of 42 trials. Each list was composed of three quartets of words that were semantically related to three common, non-presented theme words (see Appendix). The lists were counterbalanced so that each quartet appeared equally often in the primacy (‘A’), middle (‘B’), or recency (‘C’) sublist position. Item order was randomised within each sublist and occurred in a unique, randomised order for each participant.

Each of the 126 sublists was drawn from a corpus of 136 lists consisting of four words converging on a common semantic associate (Flegal, 2011). Sublists were thematically distinct lists characterised by mean backward associative strength (BAS) — a measure of how likely a given associate will be produced in response to a theme word (Arndt, 2012; Roediger et al., 2001). Lists were counterbalanced so that the mean BAS across sublists was equated across subjects ($M_{BAS} = 0.34$, $SD = 0.16$).

**Procedure**

All participants gave written informed consent at the outset of the testing session. For exploratory purposes, a battery of additional tasks including digit span (Wechsler, 1997), operation span (Unsworth, Heitz, Schrock, & Engle, 2005), and a source memory task (Drag, Bieliauskas, Kaszniak, Bohnen, & Glisky, 2009) was also given. Digit span was administered prior to the experimental session whereas the other tasks followed the experimental session. These measures will not be discussed further in the present report.

Participants were given printed instructions for the task while the experimenter read them aloud and answered any questions. Two practice trials prefaced the task to familiarise participants with its structure. Audio recording (Olympus WS-210S digital voice recorder) commenced at the beginning of the first practice trial and continued for the duration of the experiment. An experimenter remained in the testing room for the entire testing procedure.

Words were presented serially at a rate of 800 ms per word (Talmi, Grady, Goshen-Gottstein, & Moscovitch, 2005) with 50 ms between words. Participants were instructed not to read the words aloud and were removed from the final dataset if this behaviour was detected by the experimenter. Timing and presentation parameters were controlled using ePrime 2.0 software (Psychology Software Tools, Pittsburgh, PA). Stimuli appeared in a black Arial 17-point font on a silver background. Following presentation of the final list item, a row of five green “X”s appeared for 300 ms to serve both as a visual mask and as the recall cue. Participants were required to take a minimum of 30 s to freely recall the studied items aloud. When participants felt that they had recalled all words they remembered, they pressed the space bar on a standard QWERTY keyboard to advance to the next trial. Subsequent trials began either 1500 or 2000 ms following the space bar press.

**Data analysis**

Two trained coders transcribed each participant’s audio responses and coded them as either correct or incorrect. A third coder was used to resolve the rare discrepancies in coding classifications. Incorrect responses were further classified by error type within each sublist in accordance with previous work (Atkins & Reuter-Lorenz, 2008; MacDuffie, Atkins, Flegal, Clark, & Reuter-Lorenz, 2012): (1) **semantic**: responses judged by both coders as related in meaning to an item in the memory set, or the unstudied theme word itself (2) **phonological**: responses related in sound to one of the items in the memory set (3) **miscellaneous**: mispronunciations, repeats, and unintelligible utterances. Errors in the latter category are not considered further because they cannot be linked to a particular sublist but are included in Figure 1 for the interested reader. In the results that follow, correct and FM responses are expressed as proportions of the total number of responses from that sublist position. Due to the inherent family-wise nature of our comparisons (sublists A vs. B, A vs. C, B vs. C), we chose to use a Bonferroni corrected $p$-value of $p = 0.017 (03) / 3$), unless corrected otherwise for additional comparisons, as appropriate) to determine significance.

![Figure 1](image-url). (a) Enlarged view of miscellaneous responses across output positions for Experiment 1 (no distraction). Corrected by the number of responses from each position. (b) Enlarged view of miscellaneous responses across output positions for Experiment 1 (math distraction). Corrected by the number of responses from each position.
Results

Participants recalled an average of 8.18 (SD = 1.07) words per trial. Of these, an average of 7.10 (SD = 0.99) were correct responses. To look at the distribution of correct responses as a function of presentation position at encoding, we computed serial position curves (Laming, 2010; Murdock, 1962). These revealed the expected pattern of elevated memory for both early and late list positions, constituting the primacy and recency effects, respectively (Figure 2(a)). When considering output position of correct responses, items from sublist C were most likely to be the first item produced (Figure 3(a); see Hogan, 1975; Kahana et al., 2008; Laming, 2010).

Mean correct and false memory responses by sublist are presented in Table 1. Correct responses varied reliably across sublist, $F(2, 82) = 66.91, p < .001, \eta_p^2 = 0.39$. Participants correctly recalled more words from sublist C than either A, $t(41) = -10.02, p < .001, d = -1.87,^2$ or B, $t(41) = -8.65, p < .001, d = -1.7$. Numerically there were more correct responses from sublist A than B, however the difference was not reliable, $p = 0.33, d = -0.12$.

Although correct responses were dominant, we were especially interested in the distribution of false memory (FM) errors both across sublists and across FM type. Examining the relationship of FM type (semantic, phonological) by sublist (A, B, C) revealed significant main effects of sublist, $F(2, 82) = 66.91, p < .001, \eta_p^2 = 0.24$, and FM type, $F(1, 41) = 210.03, p < .001, \eta_p^2 = 0.62$. These effects arose because the rate of FM varied by sublist regardless of error type with more errors coming from the A, $t(41) = 10.02, p < .001, d = 1.87$, and B, $t(41) = 8.65, p < .001, d = 1.7$, than C sublist, and, for each sublist, there were more semantic than phonological FM errors.

Table 1. Mean (SD) number of responses per sublist, by response type and experiment.

<table>
<thead>
<tr>
<th>Response Type</th>
<th>Sublist A</th>
<th>Sublist B</th>
<th>Sublist C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>2.00 (0.41)</td>
<td>1.88 (0.49)</td>
<td>3.19 (0.46)</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>2.02 (0.45)</td>
<td>1.69 (0.42)</td>
<td>2.21 (0.56)</td>
</tr>
<tr>
<td>Semantic FM</td>
<td>0.32 (0.16)</td>
<td>0.28 (0.14)</td>
<td>0.12 (0.10)</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>0.30 (0.20)</td>
<td>0.27 (0.17)</td>
<td>0.24 (0.16)</td>
</tr>
<tr>
<td>Phonological FM</td>
<td>0.003 (0.01)</td>
<td>0.003 (0.01)</td>
<td>0.012 (0.02)</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>0.006 (0.01)</td>
<td>0.003 (0.01)</td>
<td>0.005 (0.01)</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>0.006 (0.01)</td>
<td>0.003 (0.01)</td>
<td>0.005 (0.01)</td>
</tr>
</tbody>
</table>
phonological FM, all \( p < .001 \), all \( d > 1.4 \) (Figure 4(a)). The predominance of semantic errors is unsurprising because the lists were constructed to promote these errors, and not phonological errors, which are nonetheless informative about subjects’ reliance on acoustic features of the list items (see e.g., Olszewska, Reuter-Lorenz, Munier, & Bendler, 2015).

In addition to these main effects, we also observed an interaction between sublist and FM type, \( F(2, 82) = 74.44, p < .001, \eta^2_p = 0.26 \). To understand this interaction, we further analysed FM by sublist separately for semantic and phonological FM. Follow up paired t-tests revealed that participants made more semantic FM from the A sublist, \( t(41) = 10.43, p < .001, d = 1.9 \), and B, \( t(41) = 8.97, p < .001, d = 1.74 \), than C sublist. There were numerically more semantic FMs from sublist A than B, but this difference was not reliable, \( p = 0.34, d = 0.12 \).

We also tested whether semantic FM from sublist C at the earliest output positions reliably differed from zero. This is a critical test of our claim that these FM come from a single, unitary store because it is insufficient to show that these errors are numerically present if they do not differ from zero. To establish not only that these errors were present, but that they reliably differed from zero, we computed a t-test comparing sublist C FM at each of the first four output positions against a null hypothesis of zero errors. For all four output positions, the number of sublist C FM was reliable greater than zero, \( p < .001, d > 0.54 \).

Given the relative rarity of phonological FM, we began by running t-tests to confirm that the FM rates for each sublist were reliably different from zero. Only sublist C surpassed our strict Bonferroni-corrected threshold, \( t(41) = 3.83, p < .001, d = 0.59 \). Neither sublist A, \( p = 0.07, d = 0.28 \), nor B, \( p = 0.06, d = 0.03 \), reliably differed neither from zero nor from one another, \( p = 0.9 \).

Discussion

Using the classic list-learning paradigm modified to include converging associates, this experiment demonstrates that semantic distortions can arise from all sublists, including the recency sublist. These results extend prior evidence for continuity of false memory effects across delays ranging from 3–4 s to 20-minutes or more (Flegal et al., 2010; Flegal & Reuter-Lorenz, 2014; Olszewska et al., 2015). Importantly, these new results challenge the strict view that primacy and recency positions in the serial list learning task represent retrieval from fundamentally different memory stores (Glanzer & Cunitz, 1966; Morrison, Conway, & Chein, 2014; Oztekin, Davachi, & McElree, 2010). A dual-store model might have predicted differing rates of memory performance in both veridical and FM from different sublists as we found in Experiment 1. However, dual-store models would not posit that all sublists should be susceptible to semantic distortions and thus cannot readily account for semantic FM from all sublists.

While semantic FMs are more frequently associated with primacy than recency sublists, this pattern could be explained by greater availability of perceptual features (verbatim cues) from recent sublist positions that can oppose the misleading effects of gist. If the differences in performance between primacy and recency sublists can be explained by differential access to perceptual features (Nairne, 2002) and their opposition to gist, then reducing availability of these cues should reduce veridical responses from sublist C while increasing semantic FM. To test this, in Experiment 2 we introduced a distracting arithmetic task in the retention interval prior to free recall.

Experiment 2

Method

Participants

Fifty-five University of Michigan undergraduate students took part in the experiment for course credit. All
participants were native English speakers, right-handed, and free from any reported neurological or psychological conditions. Six participants were excluded due to experimenter error and equipment malfunctions, five for low accuracy on the math distractor task (below 70%), two for audibly reading words aloud during the study phase, and one because they were a non-native English speaker. The following analyses represent data from the remaining 41 participants (24 females; \( M \) age = 18.78 years, \( SD \) = 0.99).

**Materials and procedure**

Stimuli used in Experiment 2 were identical to those in Experiment 1 except that, prior to the recall cue, a two-step math problem (based on the operation span task, (Turner & Engle, 1989), and used previously in FM studies, e.g., (Atkins & Reuter-Lorenz, 2008) appeared on the screen for 3000 ms. In this interval, participants verified whether the equation was solved correctly or incorrectly via keypress. Otherwise, task procedures were identical to those reported for Experiment 1.

**Data analysis**

Data analysis procedures were the same as in Experiment 1.

**Results and discussion**

Participants recalled an average of 7.32 (\( SD = 1.68 \)) words per trial. Of these, an average of 5.95 (\( SD = 1.11 \)) were correct responses. Once again, we computed serial position curves (Figure 2(b)) to evaluate how encoding position impacts later recall. We saw a blunted recency effect as compared to Experiment 1 (see Effects of distraction section below for quantitative between-experiment comparisons). The likelihood that sublist C items were the first items output was also blunted (Figure 3(b)).

Mean correct and FM responses by sublist are presented in Table 1. Correct responses varied reliably across sublist, \( F(2, 80) = 10.95, p < .001, \eta_p^2 = 0.06 \). Participants correctly recalled more words from sublist C, \( t(40) = -3.65, p < .001, d = -0.48 \), than A, or B, \( t(40) = -4.05, p < .001, d = -0.55 \). While the number of items correctly recalled from sublist A was numerically greater than from sublist B, this effect did not reach significance, \( p = 0.66, d = 0.05 \).

Again, we examined the distribution of FMs across sublists and FM type (semantic, phonological) and found significant main effects of sublist, \( F(2, 80) = 10.95, p < .001, \eta_p^2 = 0.03, \) and FM type, \( F(1, 40) = 151.34, p < .001, \eta_p^2 = 0.60 \). Once again, the main effect of FM type arose because, at each sublist position, there were more semantic than phonological FM, all \( p < .001, \) all \( d > 2.2 \). The main effect of sublist arose because FM rates varied by sublist position across both semantic and phonological errors such that more errors came from the A, \( t(41) = 3.65, p < .001, d = 0.48 \), and B, \( t(41) = 4.05, p < .001, d = 0.55 \), sublists than the C sublist (Figure 4(b)).

In addition to these main effects, we also observed an interaction between sublist and FM type, \( F(2, 80) = 10.76, p < .001, \eta_p^2 = 0.03 \). To understand this interaction, we looked separately at the relative rates of FM by sublist split by FM type (semantic and phonological). Follow up paired t-tests revealed that participants made more semantic FM from the \( A, t(40) = 3.70, p < .001, d = 0.48, \) and B, \( t(40) = 4.06, p < .001, d = 0.56, \) than C sublist. The numerical difference in semantic FM between A and B was not reliable, \( p = 0.60, d = -0.06 \). We additionally confirmed that sublist C FM were reliable greater than zero at all of the first four output positions, all \( p < .001, \) all \( d > 0.59 \).

Given the rarity of phonological FM, we first confirmed that the FM rates for each sublist were reliably different from zero and then ran between-sublist paired t-tests. Only sublist A surpassed our strict Bonferroni-corrected threshold, \( t(40) = 2.82, p = 0.007, d = 0.44, \) when phonological FM rates were compared against zero. Neither sublist B, \( p = 0.07, d = 0.29, \) nor C, \( p = 0.03, d = 0.35, \) reliably differed from zero based on our strict threshold nor from one another, \( p = 0.85 \).

**Effect of distraction: Experiments 1 and 2 compared**

To evaluate the effect of adding a task-filled retention interval, we conducted an experiment (1/unfilled, 2/filled) × sublist (A, B, C) ANOVA for each response type (veridical, semantic FM, phonological FM). To control for the potential inflation of significant results due to multiple comparisons, we used a strict Bonferroni-corrected \( p \)-value of \( p < .006 \) (0.05/3 sublist comparisons * 3 response types). We further note that because these experiments were run sequentially, subjects were not randomly assigned to each experiment, which constitutes a limitation of this set of analyses.

For veridical memory, there was no main effect of sublist, \( p = .042, \) however there was a main effect of experiment, \( F(2, 243) = 29.89, p < .001, \eta_p^2 = 0.20, \) that was qualified by an experiment by sublist interaction, \( F(2, 243) = 6.33, p = .002, \eta_p^2 = .05 \). Follow up t-tests revealed that veridical responses did not differ from sublist A, \( p = 0.55, d = -0.13, \) and B, \( p = 0.89, d = 0.03, \) but there were reliably more veridical responses from sublist C when there was no interfering math task immediately after the list, \( t(65.46) = 5.37, p < .001, d = 1.18. \) Thus, adding a task-filled retention interval selectively changed participants’ access to and recall from the most recent sublist.

Next, we evaluated the impact of the distracting task on memory errors. For semantic FM, the effect of experiment was not reliable, \( p = .042, \) but the effect of sublist was reliable, \( F(2, 243) = 30.91, p < .001, \eta_p^2 = 0.20, \) and qualified by a significant experiment by sublist interaction, \( F(2, 243) = 6.86, p = .001, \eta_p^2 = .05. \) Once again, the rate of semantic FM only differed between experiments for the C sublist, \( t(64.14) = -5.52, p < .001, d = -1.22, \) and not from either the A, \( p = 0.51, d = 0.15, \) nor B, \( p = 0.92, d = -0.02, \) sublists. This was due to more semantic FM from sublist C due to the inclusion of the math task.

Finally, we evaluated the impact of the distracting task on phonological FM. Here, none of the effects were reliable
(experiment, $p = 0.97$, $\eta^2_p < .001$, sublist, $p = 0.28$, $\eta^2_p = 0.01$, or their interaction, $p = 0.26$, $\eta^2_p = 0.01$). Given the rarity of phonological errors in both experiments and that the frequency of these errors by sublist was not consistently greater than zero across the two experiments, this outcome is unsurprising.

**General Discussion**

These experiments used a novel blend of the serial list learning and converging associates paradigms to investigate the prevalence of semantic distortions and susceptibility to interference across list positions. In both experiments, the majority of responses were veridical and more likely to originate from recency (sublist C) than primacy (sublist A) list positions, both of which exceeded recall from sublist B. This well-established serial position effect (Capitani et al., 1992; Craik & Levy, 1970; Glanzer & Cunitz, 1966; Laming, 2010; Murdock, 1962; Nipher, 1877; Postman & Phillips, 1965) was selectively diminished in Experiment 2 by inserting a task-filled retention interval prior to recall which reduced memory accuracy for sublist C.

The results from both experiments can be neatly explained by the differential availability of verbatim and perceptual features versus gist-based, semantic cues as a function of list position (Brainerd & Reyna, 2002; Reyna, Corbin, Weldon, & Brainerd, 2016). Verbatim, item-specific traces that code perceptual properties of studied items are short-lived and prone to interference. They are thus more likely to support retrieval from the recency sublist. Semantic or gist-based traces that code meaning are more enduring, and, when unopposed by verbatim cues, can promote false recall in addition to supporting accurate memory. To the extent that verbatim traces endure at retrieval, they oppose reliance on gist traces and support retrieval monitoring operations by providing distinctive, perceptual cues that can be monitored and utilised to reject intrusions during recall, both of which result in diminished semantic FM. In particular, other researchers (Gallo, 2004) have hypothesised that verbatim traces may also be used in recall-to-reject, distinctiveness heuristic, and other diagnostic strategies as ways to monitor memory output.

In the present data sets, varying strengths of verbatim and gist traces can explain the differences in veridical and false memories associated with different sublists and the differential impact of the filled retention interval. Verbatim traces or perceptual codes are generally more available for recently presented items than for earlier presented items leading to greater veridical recall and reduced semantic FM from sublist C, especially in Experiment 1. The distractor task prior to free recall in Experiment 2 compromised the fidelity of verbatim information and thus selectively affected sublist C leading to increased reliance on gist and more false memories associated with this sublist. Therefore, we can understand the superior memory for sublist C relative to the other two sublists, and the compromised performance for sublist C with a filled relative to an unfilled retention interval by considering the differential availability of verbatim information and how it can be used to counteract the potentially distorting effects of gist traces.

The present data are consistent with the notion that a strong gist trace, unopposed by verbatim cues, can lead to semantic FM. Barnhardt and colleagues (Barnhardt, Choi, Gerkens, & Smith, 2006) directly tested whether the strength of a trace is related to output position with the idea that semantically-related FM arise from reliance on gist traces as is also predicted by FTT. They found that presented items were more likely to be recalled earlier than non-presented but related items. In Figure 3, we show that semantic FM occurred from all output positions (and, even at the earliest four output positions, were reliably greater than zero) but that greater numbers of these FM came from later output positions. Thus, the current data provide a conceptual replication of Barnhardt and are consistent with greater availability of verbatim codes early in recall, in contrast with the enduring effects of gist. The current data additionally go beyond the suggestion from Barnhardt’s data that output position is tied to strength of the studied trace by enabling us to attribute FM to specific sublists from encoding.

The prevalence of FM in later output positions also calls attention to the shift from reliance on verbatim information early in retrieval to reliance on gist later in retrieval. Since this shift is due to the changing availability of verbatim and gist traces, shifting reliance from verbatim cues to gist can also occur by introducing a distracting task, as we show in Experiment 2. In summary, the shifting roles of verbatim and gist information is evident both in the serial position effects we observe where primacy relies more on gist and recency on verbatim cues as well as in output order where earlier output relies more on verbatim cues and later output on gist. We thank an anonymous reviewer for this point.

The present results are also compatible with the activation-monitoring account (Roediger & McDermott, 2000). Indeed, for the recency sublists in Experiment 2, decreased availability of verbatim traces to monitor and counteract the effects of gist led to increased associative errors as compared to Experiment 1 where the retention interval was unfilled. While traditionally the ideas of trace availability and monitoring have been used to explain FM in the LTM literature, here we demonstrate their additional relevance for explaining FM from the most recent studied positions with minimal delay.

Associative, or spreading, activation theories (Anderson, 1983; Howe, Wimmer, Gagnon, & Plumpton, 2009; McEvoy, Nelson, & Komatsu, 1999; Nelson, McEvoy, & Pointer, 2003; Roediger et al., 2001; Seamon, Luo, & Gallo, 1998), first espoused by Underwood (1965) as the “implicit associative response”, essentially rely on the idea that false memories are generated because at the time of learning, or retrieval,
items activate semantically-related information. Because studied lists contain multiple associatively-related items, together these activate a semantic network that also activates the non-presented, critical lure item (Seamon et al., 1998) or, in FTT terms, the gist trace. Then, at retrieval, if there is a source monitoring error (Johnson & Raye, 1981) such that subjects cannot accurately attribute an item to a specific study episode versus a covert, imagined associative response to the studied list, they may falsely remember the associated but unpresented word. Thus, activation-monitoring offers another single-store account that could potentially explain the present data.

However, one key feature of the present results is problematic for the activation-monitoring account. This account would predict greater declines in false recall than veridical recall following a filled retention interval because the associative traces of unstudied items should decay faster than traces of items that were actually studied. In Experiment 2, which included a filled retention interval, false memory rates increased along with a concomitant decrease in veridical responses from the C sublist. These patterns are more readily explained by FTT than activation-monitoring.

In both experiments, semantic FMs emerged for all sublists including the most recent sublist C, viewed traditionally as requiring recall from short-term memory. This result replicates and extends prior evidence for semantic-associative distortions in tasks considered to be canonical measures of short-term or working memory (Atkins & associative distortions in tasks considered to be canonicalally as requiring recall from short-term memory. This lists including the most recent sublist C, viewed tradition-

These patterns are more readily explained by FTT than activation-monitoring.

Thus, with respect to the current data set, TCM would predict that participants generate semantically-related FM at retrieval because encoding contexts are also retrieved during veridical recall. Retrieving the initial context can lead to generation of semantically-related (i.e., context-consistent) FM. Alternatively, retrieving the initial context can also support monitoring operations that counteract FM. The strongest evidence that encoding contexts are retrieved with item retrieval would come from observing that veridical and semantic FM from a particular sublist were clustered at output positions. The patterns of output positions in Figure 3 suggest some clustering of responses by sublist. For example, in Experiment 1 the majority of sublist C responses, regardless of their type (veridical, semantic), occur in output positions 1–4 and A responses in positions 7–12. In Experiment 2, the distribution of responses from the sublists were more evenly spread across output positions although sublist C responses still dominated at the first three output positions. Although the present pair of studies was not designed to directly adjudicate between different unitary models, they can be accommodated by various theories. An open question for future work will be to identify the points of intersection between different unitary theories or to determine if the incidence of FM can distinguish their unique predictions about the structure of memory.

In sum, both experiments provide converging evidence for the occurrence of semantic FMs across sublists, most prominently from the primacy sublist. Additionally, they indicate that the incidence of semantic FM from the recency sublist is influenced by the presence or absence of an intervening task in the retention interval. These patterns can be explained by the differential availability of gist and verbatim cues across list positions — an account that
also explains false recognition effects across delays (Flegal et al., 2010) and that obviates the need to posit separate memory stores or systems. Thus, we agree with Endel Tulving’s assessment that although you can “[...] refer to both kinds of memory as two stores, or as two systems, [...] this is] primarily for the convenience of communication, rather than as an expression of any profound belief about structural or functional separation of the two” (Tulving, 1972).

Notes

1. A post-hoc sensitivity analysis was also conducted using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). While we recognize that this is not a substitute for an a priori power analysis, it can be used to determine whether or not we had the sensitivity to detect effects should they exist. For Experiment 1 where there were 42 subjects, for our ANOVAs with within-subjects factors, assuming an alpha level of 0.05, 3 measurements (i.e., the 3 sublists), and a default 0.5 correlation between the measures and a non-sphericity correction of 1, the critical F value would be 3.12 with an effect size of 0.25. In the corresponding follow up t-tests (difference between matched pairs which is appropriate given that all comparisons were done within subjects), the critical value would be 2.02 with an effect size dz of 0.57. For Experiment 2 where there were 41 subjects, the critical F would be 3.11 with an effect size of 0.26 and the critical t would be 2.02 and an effect size of 0.58. Thus, it seems that we had adequate sensitivity to detect our observed effects.

2. Cohen’s d effect sizes for within-subject paired samples t-tests were computed correcting for correlation between means, as suggested by (Dunlap, Cortina, Vaslow, & Burke, 1996).

3. Restricting semantically-related FM to recall of lure items did not change distribution of FM nor the pattern of results. Thus, we present lure and other semantically related FM errors combined.

4. Given that phonological errors were rare, we repeated our analyses with an arcsine transformation (see (Howell, 1997; MacDufﬁe et al., 2012)) to correct for non-normality in the distribution. Performing this transform prior to running standard parametric statistics did not change our interpretation of the rate of phonological FM for either experiment alone nor in our between experiment comparisons. We also note that because the frequency of these errors by sublist was not consistently greater than zero across the two experiments, this limits the conclusions that can be drawn based on comparison of these error rates, however, we include them for completeness.

5. FTT does not take a strong stance on the fundamental organization of the underlying memory stores although it does favor a two-stage Markov model for modelling recall that separates recollection- or verbatim-based responses from familiarity- or gist-based responses (Gomes, Brainerd, Nakamura, & Reyna, 2014).

Acknowledgments

The authors wish to thank Tom Bergman, Tamara Brodsky, Ashley Brooks, Jackie Dobson, Autumn Gear, Rachel German, Katherine Khatibi, Michael Kinning, Nina Massad, and Julia Tattan for their help with data collection and coding.

Disclosure statement

No potential conflict of interest was reported by the authors.

References


ORCID

Halle R. Dimsdale-Zucker http://orcid.org/0000-0002-8611-111X
### Appendix. The forty-two 12-item lists used in Experiments 1 & 2, listed alphabetically by sublist theme

<table>
<thead>
<tr>
<th>ACT-SHOVEL- SLEEP</th>
<th>ADJECTIVE-SNAKE-SOFT</th>
<th>ADULT-BLOOD-SHIRT</th>
<th>AGAIN-BARBECUE-FRIEND</th>
</tr>
</thead>
<tbody>
<tr>
<td>perform</td>
<td>adverb</td>
<td>mature</td>
<td>repeat</td>
</tr>
<tr>
<td>portray</td>
<td>pronoun</td>
<td>responsible</td>
<td>never</td>
</tr>
<tr>
<td>drama</td>
<td>noun</td>
<td>grown</td>
<td>twice</td>
</tr>
<tr>
<td>pretend</td>
<td>verb</td>
<td>kid</td>
<td>stop</td>
</tr>
<tr>
<td>dig</td>
<td>slither</td>
<td>plasma</td>
<td>cookout</td>
</tr>
<tr>
<td>spade</td>
<td>serpent</td>
<td>donor</td>
<td>grill</td>
</tr>
<tr>
<td>pail</td>
<td>reptile</td>
<td>vein</td>
<td>ribs</td>
</tr>
<tr>
<td>rake</td>
<td>rattle</td>
<td>vampire</td>
<td>sauce</td>
</tr>
<tr>
<td>nap</td>
<td>hard</td>
<td>blouse</td>
<td>pal</td>
</tr>
<tr>
<td>doze</td>
<td>loud</td>
<td>sleeves</td>
<td>buddy</td>
</tr>
<tr>
<td>bed</td>
<td>tender</td>
<td>collar</td>
<td>companion</td>
</tr>
<tr>
<td>awake</td>
<td>fluffy</td>
<td>shorts</td>
<td>neighbor</td>
</tr>
<tr>
<td>ALONE-ARMY-FOG</td>
<td>ANGER-SNEEZE-STOMACH</td>
<td>ANNOY-CITY-ROCK</td>
<td>ANSWER-BLACK-CABBAGE</td>
</tr>
<tr>
<td>isolated</td>
<td>rage</td>
<td>bother</td>
<td>question</td>
</tr>
<tr>
<td>solo</td>
<td>mad</td>
<td>aggravate</td>
<td>reply</td>
</tr>
<tr>
<td>lonely</td>
<td>enrage</td>
<td>irritate</td>
<td>response</td>
</tr>
<tr>
<td>one</td>
<td>fury</td>
<td>disturb</td>
<td>solution</td>
</tr>
<tr>
<td>navy</td>
<td>allergy</td>
<td>metroplis</td>
<td>white</td>
</tr>
<tr>
<td>soldier</td>
<td>cough</td>
<td>town</td>
<td>gray</td>
</tr>
<tr>
<td>infantry</td>
<td>handkerchief</td>
<td>urban</td>
<td>brown</td>
</tr>
<tr>
<td>marines</td>
<td>tissue</td>
<td>suburb</td>
<td>coal</td>
</tr>
<tr>
<td>mist</td>
<td>abdomen</td>
<td>Boulder</td>
<td>patch</td>
</tr>
<tr>
<td>haze</td>
<td>belly</td>
<td>stone</td>
<td>sauerkraut</td>
</tr>
<tr>
<td>smog</td>
<td>intestine</td>
<td>solid</td>
<td>slaw</td>
</tr>
<tr>
<td>unclear</td>
<td>ulcer</td>
<td>roll</td>
<td>lettuce</td>
</tr>
<tr>
<td>ARGUE-SINK-SPIDER</td>
<td>ATOM-BROOM-LIE</td>
<td>AUTHOR-LOSE-NEEDLE</td>
<td>BABY-BASKET-THIEF</td>
</tr>
<tr>
<td>debate</td>
<td>molecule</td>
<td>writer</td>
<td>crib</td>
</tr>
<tr>
<td>disagree</td>
<td>nucleus</td>
<td>poet</td>
<td>infant</td>
</tr>
<tr>
<td>complain</td>
<td>neutron</td>
<td>editor</td>
<td>diaper</td>
</tr>
<tr>
<td>agree</td>
<td>proton</td>
<td>publisher</td>
<td>carriage</td>
</tr>
<tr>
<td>drain</td>
<td>dustpan</td>
<td>win</td>
<td>wicker</td>
</tr>
<tr>
<td>float</td>
<td>sweep</td>
<td>find</td>
<td>picnic</td>
</tr>
<tr>
<td>faucet</td>
<td>mop</td>
<td>gain</td>
<td>waste</td>
</tr>
<tr>
<td>bathroom</td>
<td>witch</td>
<td>defeat</td>
<td>laundry</td>
</tr>
<tr>
<td>web</td>
<td>fib</td>
<td>thread</td>
<td>crook</td>
</tr>
<tr>
<td>tarantula</td>
<td>untruthful</td>
<td>syringe</td>
<td>robber</td>
</tr>
<tr>
<td>arachnid</td>
<td>deception</td>
<td>haystack</td>
<td>burglar</td>
</tr>
<tr>
<td>creepy</td>
<td>dishonest</td>
<td>injection</td>
<td>bandit</td>
</tr>
<tr>
<td>BACK-GAS-HELP</td>
<td>BAKE-KING-ROOF</td>
<td>BEAUTIFUL-CORN-SLOW</td>
<td>BEE-BUY-WISH</td>
</tr>
<tr>
<td>front</td>
<td>broil</td>
<td>gorgeous</td>
<td>hive</td>
</tr>
<tr>
<td>spine</td>
<td>oven</td>
<td>lovely</td>
<td>bumble</td>
</tr>
<tr>
<td>behind</td>
<td>cook</td>
<td>pretty</td>
<td>sting</td>
</tr>
<tr>
<td>forward</td>
<td>cake</td>
<td>ugly</td>
<td>buzz</td>
</tr>
<tr>
<td>fuel</td>
<td>throne</td>
<td>cob</td>
<td>purchase</td>
</tr>
<tr>
<td>petroleum</td>
<td>queen</td>
<td>husk</td>
<td>sell</td>
</tr>
<tr>
<td>station</td>
<td>crown</td>
<td>flame</td>
<td>store</td>
</tr>
<tr>
<td>oil</td>
<td>reign</td>
<td>field</td>
<td>spend</td>
</tr>
<tr>
<td>assist</td>
<td>shingle</td>
<td>fast</td>
<td>hope</td>
</tr>
<tr>
<td>aid</td>
<td>ceiling</td>
<td>snail</td>
<td>want</td>
</tr>
<tr>
<td>emergency</td>
<td>tar</td>
<td>turtle</td>
<td>desire</td>
</tr>
<tr>
<td>wanted</td>
<td>tin</td>
<td>sluggish</td>
<td>dream</td>
</tr>
<tr>
<td>BOOM-FISH-TEETH</td>
<td>BOX-CHURCH-TRIP</td>
<td>BREAD-MOUNTAIN-REN</td>
<td>BUILDING-MORNING-</td>
</tr>
<tr>
<td>sonic</td>
<td>cardboard</td>
<td>rye</td>
<td>structure</td>
</tr>
<tr>
<td>bang</td>
<td>carton</td>
<td>loaf</td>
<td>blocks</td>
</tr>
<tr>
<td>explosion</td>
<td>storage</td>
<td>butter</td>
<td>construction</td>
</tr>
<tr>
<td>bomb</td>
<td>container</td>
<td>toast</td>
<td>empire</td>
</tr>
<tr>
<td>trout</td>
<td>cathedral</td>
<td>climber</td>
<td>early</td>
</tr>
<tr>
<td>cod</td>
<td>steeple</td>
<td>hill</td>
<td>dawn</td>
</tr>
<tr>
<td>scales</td>
<td>temple</td>
<td>climb</td>
<td>dew</td>
</tr>
<tr>
<td>shrimp</td>
<td>preacher</td>
<td>molehill</td>
<td>evening</td>
</tr>
<tr>
<td>gums</td>
<td>journey</td>
<td>own</td>
<td>aroma</td>
</tr>
<tr>
<td>braces</td>
<td>vacation</td>
<td>lease</td>
<td>scent</td>
</tr>
</tbody>
</table>

(Continued)
mouth

travel

apartment

whiff

tongue

baggage

monthly

stench

BUTTERFLY-COLD-PAN

CARPET-MAN-WINDOW

CHAIN-FAIL-JOB

CHAIR-PARK-ROPE

cocoon

rug

link

moth

floor

whip

insect

magic

necklace

wing

red

bicycle

hot

woman

flunk

shiver

arctic

handsome

try

frigid

male

succeed

skillet

pane

occupation

pot

sill

career

dish

shutter

twine

CHAOS-FOOT-FRUIT

CHEESE-LION-MUSCLE

CLAM-DESTROY-FINISH

COPY-EGGS-FRAGILE

havoc

cheddar

chowder

anarchy

swiss

oyster

hectic

cracker

shell

confusion

mouse

mussel

toe

roar

demolish

inch

tanner

ruin

ankle

tiger

annihilate

shoe

mane

create

kiwi

flex

done

citrus

weights

breakable

pear

strength

fail

berry

ton

glass

CUP-PEN-PULL

DANCE-MARRY-MOVIE

DOCTOR-DRY-GIRL

FLAG-RIVER-SMOKE

saucer

ballet

physical

measuring

ballerina

nurse

mug

song

stethoscope

goblet

aerobics

surgeon

quill

wed

towel

pencil

engage

desert

marker

single

moist

write

hitch

thirst

tug

cinema

boy

push

film

dolls

drag

theater

female

stretch

popcorn

dress

FLOWER-MAP-WHOLE

FOREVER-HEestead-ZUCKER ET AL.

FUNNY-HEALTH-PIG

GIVE-HIGH-NUT

tulip

eternity

hilarious

petals

infinity

comedian

daisy

always

humor

vase

endless

clown

atlas

ghoul

sickness

chart

goblin

body

direction

phantom

wealth

world

spook

ill

half

liberty

holl

part

courts

pork

piece

truth

sow

all

lawyer

st

HOLD-PRESENT-SQUARE

HORSE-Piano-RUBBER

LETTER-RING-TRASH

MANy-RAIN-ROUGH

grasp

saddle

envelope

grip

pony

stamp

keep

gallup

mailbox

carry

colt

mail

gift

keyboard

diamond

past

organ

bell

absent

guitar

jewelry

future

ivy

phone

circle

foam

garbage

triangle

latex

garbage

round

galoshes

debris

shape

tire
<table>
<thead>
<tr>
<th>MATH-SAFE-STRESS</th>
<th>SMART-SPICE-VOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>arithmetic</td>
<td>intelligent</td>
</tr>
<tr>
<td>calculus</td>
<td>genius</td>
</tr>
<tr>
<td>algebra</td>
<td>wise</td>
</tr>
<tr>
<td>equation</td>
<td>knowledge</td>
</tr>
<tr>
<td>vault</td>
<td>oregano</td>
</tr>
<tr>
<td>secure</td>
<td>herb</td>
</tr>
<tr>
<td>guard</td>
<td>cinnamon</td>
</tr>
<tr>
<td>lock</td>
<td>seasoning</td>
</tr>
<tr>
<td>tension</td>
<td>ballot</td>
</tr>
<tr>
<td>pressure</td>
<td>election</td>
</tr>
<tr>
<td>anxiety</td>
<td>register</td>
</tr>
<tr>
<td>strain</td>
<td>campaign</td>
</tr>
</tbody>
</table>