

Semantic and phonological contributions to short-term repetition and long-term cued sentence recall

Jed A. Meltzer^{1,2}  · Nathan S. Rose¹ · Tiffany Deschamps¹ · Rosie C. Leigh¹ · Lilia Panamsky¹ · Alexandra Silberberg¹ · Noushin Madani¹ · Kira A. Links¹

© Psychonomic Society, Inc. 2015

Abstract The function of verbal short-term memory is supported not only by the phonological loop, but also by semantic resources that may operate on both short and long time scales. Elucidation of the neural underpinnings of these mechanisms requires effective behavioral manipulations that can selectively engage them. We developed a novel cued sentence recall paradigm to assess the effects of two factors on sentence recall accuracy at short-term and long-term stages. Participants initially repeated auditory sentences immediately following a 14-s retention period. After this task was complete, long-term memory for each sentence was probed by a two-word recall cue. The sentences were either concrete (high imageability) or abstract (low imageability), and the initial 14-s retention period was filled with either an undemanding finger-tapping task or a more engaging articulatory suppression task (Exp. 1, counting backward by threes; Exp. 2, repeating a four-syllable nonword). Recall was always better for the concrete sentences. Articulatory suppression reduced accuracy in short-term recall, especially for abstract sentences, but the sentences initially recalled following articulatory suppression were retained *better* at the subsequent cued-recall test, suggesting that the engagement of semantic mechanisms for short-term retention promoted encoding of the sentence meaning into long-term memory. These results provide a basis for using sentence imageability and subsequent memory

performance as probes of semantic engagement in short-term memory for sentences.

Keywords Short-term memory · Long-term memory · Cued recall · Semantic · Phonological · Working memory · Sentence repetition

Short-term memory (STM) comprises multiple mechanisms that maintain different kinds of information. Since the proposal of the influential multicomponent model of working memory (Baddeley, 2003; Baddeley & Hitch, 1974), theorists have distinguished between the phonological loop, responsible for subvocal rehearsal of verbal information, and the visuospatial sketchpad, responsible for maintenance of visual information over short delays. Subsequent research has suggested further fractionation of verbal STM into other resources beyond the phonological loop. Although phonological STM (pSTM) is critical for the maintenance of arbitrary information such as digit strings and nonwords, semantic mechanisms can complement pSTM to support the maintenance of more meaningful information. For example, neuropsychological patients with impairments of the phonological loop fail on short-term recall of arbitrary lists, but they can often produce reasonable paraphrases of meaningful sentences (Baldo, Klosternann, & Dronkers, 2008; Butterworth, Shallice, & Watson, 1990; McCarthy & Warrington, 1984, 1987). Healthy participants can perfectly recall sentences that greatly exceed their estimated span for word lists in length (Brenner, 1940; Miller & Selfridge, 1950). Although chunking and prediction may play roles in sentence recall, semantic factors also affect the recall of word lists, with superior recall for words versus nonwords (Hulme, Maughan, & Brown, 1991), for high- versus low-frequency words (Hulme et al., 1997), and for words of higher imageability (Bourassa & Besner, 1994).

✉ Jed A. Meltzer
jmeltzer@research.baycrest.org

¹ Baycrest Centre, Rotman Research Institute, 3560 Bathurst Street, Toronto, Ontario M6A 2E1, Canada

² Departments of Psychology and Speech-Language Pathology, University of Toronto, 27 King's College Circle, Toronto, Ontario M5S, Canada

Although experimental tasks can be designed to stress phonological or semantic maintenance, both kinds of information likely contribute to verbal STM in everyday life. A classic task involving both is sentence repetition. Although early viewpoints tended to view phonological mechanisms as the primary contributor to sentence repetition (e.g., Clark & Clark, 1977), a radically different view was supported by Potter and Lombardi (1990), who proposed that short-term sentence recall depends mainly on regeneration from a conceptual code. This was evidenced by participants' tendency to make semantic substitutions in their sentence repetitions when semantically related lure words were contained within word lists used in secondary tasks performed immediately before or after the sentence presentation. Subsequent studies using similar paradigms have shown that both semantically and phonologically related lure words tend to intrude in sentence recall attempts, supporting a more mixed viewpoint in which both semantic and phonological codes contribute to sentence recall performance (Alloway, 2007; Rummer & Engelkamp, 2001, 2003; Schweppe, Rummer, Bormann, & Martin, 2011).

The parallel engagement of phonological and semantic mechanisms suggests that the two processes may interact with each other. One key question is whether phonological and semantic maintenance are competitive with each other. In the present study, we hypothesized that suppressing pSTM during a short-term sentence repetition task would encourage participants to increase their employment of semantic maintenance strategies, and that this manipulation would have consequences for the long-term retention of sentence content when it was tested in a subsequent cued-recall test. The logic of the present experiment rests on a close link between semantic processing and the encoding of verbal information into long-term memory (LTM). It is well known that making semantic decisions about verbal stimuli promotes their encoding into LTM, relative to perceptual or phonological decisions, a finding known as the "levels-of-processing" effect (Craik & Lockhart, 1972). The close link between semantic processing and LTM encoding, along with other evidence, has led some theorists to propose that a single mechanism can account for both short-term and long-term retention of verbal information, and that semantic STM represents the temporary activation of representations stored in LTM (Cameron, Haarmann, Grafman, & Ruchkin, 2005; Ruchkin, Grafman, Cameron, & Berndt, 2003). This viewpoint is implied by Baddeley's (2000) use of the term "episodic buffer" to refer to the short-term storage of verbal information by mechanisms other than the phonological loop. Such information may be temporarily stored via LTM mechanisms but ultimately fails to be consolidated for longer-term retention, such that the episodic buffer serves as both an STM store and a gateway into LTM. Other theorists have posited the existence of a dedicated storage buffer independent of both pSTM and LTM, labeled variously as "semantic STM" (R. C. Martin & He, 2004) or "conceptual STM" (cSTM; Potter & Lombardi, 1990). Experiments have

identified specific effects that seem to call for a buffer independent of LTM (Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Davelaar, Haarmann, Goshen-Gottstein, & Usher, 2006; Haarmann, Davelaar, & Usher, 2003; Haarmann & Usher, 2001; Shivde & Anderson, 2011), although a contribution of LTM to short-term sentence recall is not ruled out by such findings.

In the present study, we do not aim to distinguish between dual- and single-mechanism accounts of semantic maintenance, but rather seek to clarify the relative contributions of pSTM and semantic mechanisms toward the encoding of sentence content into LTM. We evaluated memory for sentence content by having participants recall the same sentences twice. In Task 1 (short-term repetition), participants heard a lengthy sentence at the beginning of each trial and were asked to attempt to repeat it verbatim after a 14-s delay. During the delay, they were asked to either tap their fingers or perform a phonologically demanding distractor task, with both actions being externally paced by a visual cue. The more demanding distractor task involved counting backward by threes in Experiment 1, and repeating a nonword in Experiment 2. The distractor conditions in both experiments involved articulatory suppression (AS) of pSTM and were expected to produce poorer performance in short-term repetition. For effective short-term recall of sentences following AS, participants must rely on alternative mechanisms either to maintain the sentence content in a buffer during the delay or to recall the sentence from LTM after the delay is over. We hypothesized that this engagement of alternative mechanisms would have consequences for the degree to which the sentence content was ultimately retained in LTM, as tested on a second task.

Task 2 (long-term cued recall) involved a somewhat novel paradigm. On each trial, participants were presented with two words (the subject and main verb) from a sentence they had previously encountered in Task 1. They were asked to recall as much of the sentence as possible on the basis of this two-word cue. Sentences were scored according to raw recall performance (how many words from the original sentence were recalled), but also, critically, for conditional recall (the proportion of words recalled in Task 2 relative to the number recalled for the same sentence on Task 1). The conditional recall score specifically assessed the degree to which the sentence was forgotten following its initial short-term recall, before it was cued in the second task. The raw recall score reflects forgetting (or failure to encode) at two stages—during the initial delay in Task 1, and between the two tasks.

On the basis of the principle of levels of processing, we predicted in the present experiments that short-term repetition of sentences under conditions of AS would result in higher conditional recall scores, which would indicate less forgetting between repetition on Task 1 and cued recall on Task 2. Although we expected that AS would produce lower raw recall scores than the tapping condition on both tasks, due to the

words forgotten during the initial filled delay period, we expected participants to forget fewer words between the two recall attempts when the first attempt was performed following AS. This prediction stems from the expected use of semantically based STM and LTM strategies during the initial recall attempt, which should promote encoding into LTM. This prediction can be contrasted with an opposite prediction that would view pSTM-based repetition as being more beneficial for encoding information into LTM. Even acknowledging the existence of levels-of-processing effects, one might predict that the benefit of rehearsing a sentence in the phonological loop without interference might outweigh the advantage afforded by being forced to recall the sentence using semantic resources, especially if semantic processing happens inadvertently in the course of rehearsal. This would be in line with numerous studies that have demonstrated an LTM advantage for verbal stimuli that are rehearsed for longer periods of time (e.g., Aldridge & Crisp, 1982; Dark & Loftus, 1976; Rundus, 1977). Thus, a levels-of-processing perspective may be contrasted with a more intuitive “rehearsal advantage” perspective.

We hypothesized that for the AS condition, the increased effort exerted to maintain the sentence in cSTM, or to recall it from LTM after the delay, might engage semantic mechanisms that would actually promote the encoding of the sentence into LTM. During the easier finger-tapping condition, participants freely engaged in subvocal rehearsal, which yields superior short-term recall but may promote shallower encoding, and therefore more forgetting of the sentence content between the two tasks. This assumes that AS is more disruptive of pSTM than is finger tapping, such that the finger-tapping condition afforded the participants the luxury of rehearsing without much interference. To assure that this was the case, we used a simple version in which participants only had to tap one finger, synchronized to periodic visual cues on the screen. This condition matched the AS conditions in timing and the amount of visual information, but was cognitively very undemanding. Although finger-tapping tasks are occasionally used to cause dual-task interference with memory processing (see, e.g., Kane & Engle, 2000), such tasks typically require endogenous timing and complex sequences. The assertion that finger tapping minimally interfered with pSTM is supported by the high performance of participants on short-term repetition (Task 1) in the tapping condition (as compared to the much poorer repetition following AS; see the results), and also by the responses to a debriefing questionnaire administered to the participants on the strategies used.¹

In contrast to the intuitive “rehearsal advantage account,” we hypothesized here that AS would have different effects on

short-term and long-term retention of sentences. Although sentence recall had not to our knowledge previously been tested at both short-term and long-term stages, as in this experiment, studies have shown that increased demands for semantic processing during short-term retention result in improved subsequent memory for single words as assessed by recognition (Rose & Craik, 2012; Rose, Myerson, Roediger, & Hale, 2010) or free recall (McCabe, 2008; Rose, 2013; Rose, Buchsbaum, & Craik, 2014). On the basis of these findings, we expected that the AS conditions would bring about increased processing of sentence content in cSTM or LTM, resulting in less forgetting (better conditional recall) of sentences initially recalled after AS.

In addition to testing the effects of distraction on short-term repetition and long-term cued recall, the experiments incorporated a second manipulation intended to further probe the selective engagement of semantic mechanisms in short-term retention. Sentences were designed in two conditions: abstract and concrete. *Abstract* sentences were designed to be relatively devoid of sensory information, whereas *concrete* sentences described scenes that were rich in visual information. The greater imageability of the concrete sentences would support semantic memory for their content, making them less dependent on rote rehearsal for successful recall. We therefore predicted better recall for the concrete sentences in both short-term and long-term recall tasks, in line with prior experiments (Paivio, Clark, & Khan, 1988). Furthermore, we predicted an interaction between imageability and distraction condition. Studies on single-word recall have shown that concrete words benefit more from semantic relative to phonological processing in incidental encoding than abstract words do, due to their stronger links to amodal semantic representations (D’Agostino, O’Neill, & Paivio, 1977). In the present experiments, we expected to find distinct interaction effects in the two tasks. For short-term repetition, we expected that participants would be more dependent on rote rehearsal to support the maintenance of sentence content for abstract sentences, since the lack of imageability would make the semantic support weaker. Therefore, abstract sentences should suffer more degradation in performance from AS than would concrete sentences. But for conditional performance on long-term cued recall, we expected AS to have a greater *positive* impact on abstract sentences (i.e., we should see less forgetting for abstract sentences initially recalled after AS), due to participants having been forced to process these sentences semantically when they otherwise would have relied chiefly on pSTM. For concrete sentences, we also expected a beneficial effect of AS on conditional recall, but it might be smaller than the effect on abstract sentences, since concrete sentences already benefit from stronger semantic support even in the absence of AS, such as in the finger-tapping condition.

In addition to quantifying accuracy, we also analyzed sentence recall performance for the frequencies of distinct kinds of errors. Recall of sentences based on meaning rather than a

¹ The participants in Experiment 1 were given a brief questionnaire asking about their strategies used (data are not shown). One question asked them to rate from 1 to 10 the degree to which counting backward interfered with their ability to rehearse the sound of the sentence, relative to trials not involving counting (the finger-tapping trials). The mean response was 8.08, indicating a high subjective degree of interference.

phonological trace predicts a higher frequency of certain kinds of errors, particularly semantic substitutions. We observed differences in the patterns of errors induced by the two experimental factors in the stages of short-term repetition and long-term cued recall, further elucidating the complementary interactions between the phonological and semantic memory systems.

Method

Materials: sentence construction

All sentences were written by the authors to be suitable for both tasks used in the experiment: short-term repetition and long-term cued recall. For the cued-recall paradigm, the subject and verb of the sentence's main clause served as the retrieval cue. Therefore, sentences were deliberately constructed to put these two words at the beginning of the sentence, slightly constraining the syntactic structures used. The sentences were intended to fall into two conditions, "abstract" and "concrete," although independent raters were subsequently used to verify the distinction and select the final stimulus set. All of the concrete sentences contained rich visual imagery. Traditionally, "concreteness" is defined as the extent to which a concept can be experienced through the senses, whereas "imageability" is specific to the visual modality (Richardson, 1975). Although some authors dissociate these concepts, they are highly correlated, and subjective concreteness ratings are dominated by visual experience (Brysbaert, Warriner, & Kuperman, 2014). Thus, in the present experiments, we, like many other researchers (e.g., Reilly & Kean, 2007) considered the terms "concrete" and "imageable" to be interchangeable.

We composed 198 sentences ranging from 10 to 16 words in length (median = 13), corresponding to previous estimates of sentence memory span (Baddeley, Vallar, & Wilson, 1987; Brener, 1940). To confirm our intuitions about the distinction between abstract and concrete sentences, we recruited three raters naïve to the purpose of the experiment. The raters were instructed to rate the sentences according to imageability, on a scale from 1 (*least*) to 5 (*most*). See Appendix A for the instructions given to the raters. Preliminary analysis of the average ratings revealed a bimodal distribution, with abstract sentences consistently being rated 1–2 and concrete sentences being rated 3–5. We next eliminated all sentences with an average rating falling between 2 and 3, as well as all sentences with a standard deviation of the ratings greater than 1.145, such that ratings of [3, 5, 5] were acceptable, but [2, 5, 5] were not. Additional sentences were excluded on the basis of perceived semantic overlap and other subjective criteria, leaving 162 sentences that were considered acceptable for the experiment. Of these, the concrete sentences had a mean imageability rating of 4.54, and the abstract sentences of 1.17 [$t(115) = 44.3, p < 10^{-15}$].

Of the remaining 162 sentences, we then selected four sets of 25 sentences, two sets each for concrete and abstract. The sets were matched on a variety of quantitative criteria computed using readily available tools from the computational-linguistics literature (details can be found in Appendix B). The full set of 100 selected sentences is given in Appendix D. The matched sets were randomly assigned to the counting and tapping conditions, counterbalanced across participants. Sentences were recorded by a theatrically trained female speaker at a natural speaking rate (175 words, or 281 syllables, per minute).

Experiment 1: short-term repetition and long-term cued recall with an arithmetic distraction task

Participants

Twenty right-handed young adult participants (mean age = 21.9, $SD = 2.87$; 12 females, eight males) were recruited from local universities. All had spoken English fluently from age 5 or earlier and reported no history of neurological conditions or speech/hearing difficulties. All procedures were approved by the Research Ethics Board of Baycrest Hospital, and participants were compensated financially.

Procedure

Participants were tested individually in a private room, seated in front of a computer monitor. The experiment was implemented using the Presentation software (Neurobehavioral Systems, Albany, CA, USA). The experimenter was present but did not interact with the participants except to instruct them on the task and verify compliance. Auditory sentence stimuli were presented through headphones adjusted by the participant to a comfortable listening level with an attached microphone. All verbal responses were recorded for subsequent analysis.

Task 1: short-term repetition Participants completed two brief practice sessions before the experiment. First, they practiced the counting task (described below) without the concurrent sentence recall task. All participants were able to comfortably count out loud, paced with the visual cues, after five to ten practice trials. Next, the participants practiced four trials of the full task, to ensure that they understood the instructions.

This experiment featured a 2×2 factorial design, for a total of four conditions. One factor was Sentence Type, abstract or concrete (see below), but this factor was not disclosed to the participants. The second factor was the Delay Period Task, either counting (a task involving AS) or finger tapping (a nondemanding control task). The trial structure is diagrammed in Fig. 1. The counting task involved counting backward by threes from a random number between 50 and 150 that was

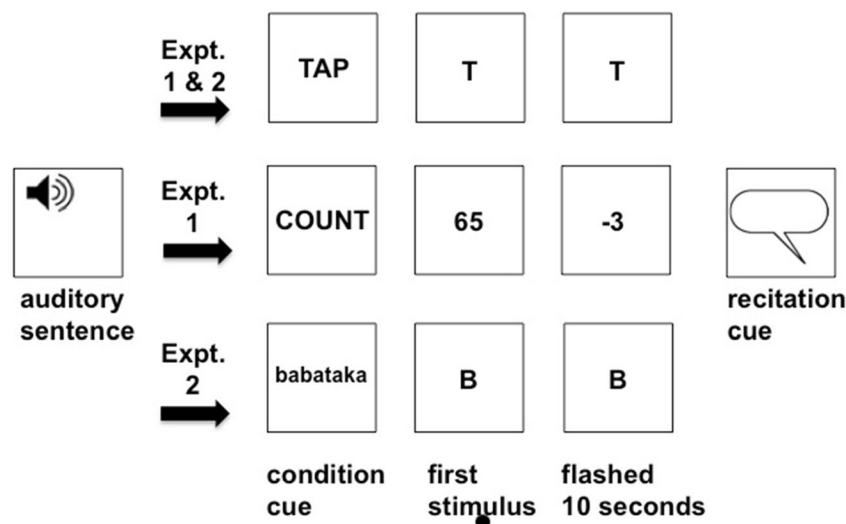


Fig. 1 Schematic of the trial structure for the short-term recall task. The top line shows the structure of tapping trials in both experiments. The middle and bottom lines show the articulatory suppression conditions in

Experiment 1 (counting aloud backward by threes) and Experiment 2 (nonword repetition), respectively

presented on the screen. At the start of each trial, participants heard a sentence, followed by a 2-s delay. On counting trials, a visual text cue then appeared with the initial random number—for example, “COUNT FROM 115.” This cue was displayed for 1,000 ms, followed by a 1,000-ms delay. Next, the visual cue “-3” appeared regularly (500 ms on, 1,000 ms off) a total of seven times over 10.5 s. Participants were instructed to say the next number in the series out loud, paced to the cue. For the tapping task, after an initial “TAP” cue (1,000 ms on, 1,000 ms off), the letter “T” appeared regularly on the screen 14 times (500 ms on, 250 ms off) over 10.5 s. Participants were instructed to tap their fingers on the table in front of them paced to the cue. After the delay period task (tapping or counting), an empty word balloon appeared on the monitor, cueing the participant to recall the sentence that he or she had just heard. The instruction “Press space bar when finished” appeared below. Participants were instructed to attempt to reproduce the sentence verbatim. When finished, they pressed the space bar, at which point the visual cue disappeared, a 2-s delay occurred, and the next trial began. A maximum of 20 s was allowed for each verbal response, after which the visual cue would disappear and the next trial would begin, but participants almost always pressed the space bar well before the time limit. After every 20 trials, they were given the opportunity to rest briefly. In total, 100 trials were presented—25 in each condition, intermixed in a random order. Sentences were presented in the same fixed random order for each participant, but the assignments of sentences to the tapping and counting conditions were counterbalanced.

Task 2: long-term cued recall Immediately after completion of Task 1, participants were told that they would undergo a second, separate memory test, consisting of cued sentence

recall. Participants were then instructed on the second task and given another practice session of four trials. The conditions for the cued-recall test involved the same 2×2 design as the immediate-recall task, but the task procedure was exactly the same for all trials (i.e., there was no distinction between “tap” and “count” trials in the task demands for cued recall—only differences in the conditions under which the sentences had *previously* been recalled). The sentences were the same 100 used for Task 1, presented in randomized order. They were either abstract or concrete, and had previously been recalled in Task 1 following tapping or counting. At the start of each trial, a word balloon appeared on the screen containing two words as a retrieval cue. Again, the instruction “Press space bar when finished” appeared at the bottom of the screen. Participants were asked to attempt to recall verbatim the sentence from Task 1 that had contained the two cue words as subject and verb. Again, a maximum of 20 s was given for each response, but the experiment was otherwise self-paced.

Experiment 2: short-term repetition and long-term cued recall with simple AS

In Experiment 1, we had used a backward-counting task to disrupt phonological rehearsal during the delay period in the short-term recall task. This task is somewhat more difficult and attentionally demanding than more traditional articulatory distraction tasks, which typically involved repetition of a single word or phrase. We chose this task in order to more completely disrupt processes that might aid in retention of the phonological form of sentences over the delay period, including articulatory rehearsal and attentional refreshing or covert retrieval (Rose et al., 2014). However, we were concerned that the results may have been influenced by the suppression of additional cognitive

processes, beyond phonological rehearsal, so we chose to replicate the experiment using a more traditional AS task as the more challenging distractor condition. The experimental design was otherwise identical to that of Experiment 1 (see Fig. 1). Fifteen participants (mean age = 21.5, $SD = 4.53$; nine females, six males) were recruited, meeting the same criteria as in Experiment 1. The finger-tapping condition was identical. On AS trials for Task 1 (short-term repetition), the nonword “BABATAKA” appeared on the screen for 1,000 ms, followed by a 1,000-ms delay. Next, the visual cue “B” appeared regularly (500 ms on, 1,000 ms off) a total of seven times over 10.5 s. Participants were instructed to pronounce “BABATAKA” out loud, paced to the cue. Thus, the visual pacing was the same as in the first experiment, but participants only had to repeat a simple four-syllable sequence instead of performing mental arithmetic. Task 2 (long-term cued recall) was identical to that in Experiment 1.

Analysis of recall accuracy

Verbatim Full details of the analysis procedure are given in Appendix C, with examples. Verbal responses were manually transcribed. The procedures for scoring short-term repetition and long-term cued-recall trials was the same. We computed a strict “verbatim” score on the basis of recall of exact word forms, and also a more liberal “gist” score accounting for paraphrases of the sentence content.

The primary measure of interest was the proportion of words in the sentence recalled correctly verbatim, ranging from 0 to 1. For a word to be scored as correctly recalled, it had to be identical to the target word, including grammatical inflections such as tense and plurality. Credit was given for open-class words that were recalled correctly, even if their serial order was changed in the response. The transcribed response was then compared with the target sentence to produce a “condensed” transcription consisting only of the correctly recalled words. For both immediate and delayed recall, the score on each trial was computed as follows:

$$\text{Recall} = \# \text{words in condensed transcription} / \# \text{words in original target sentence}.$$

For the long-term cued-recall task, the raw accuracy was not of primary interest. Failure to recall words in this task could result from two stages of forgetting: over the initial filled delay (also influenced by failure to encode initially), and during the intervening time between the short-term repetition trial and the subsequent cued-recall trial. In general, for long-term recall, participants would not be expected to correctly recall words of a sentence that they did not recall upon short-term repetition of the same sentence. Although such “reminiscences” do occur occasionally (Erdelyi, 2010), words that are forgotten during the immediate delay interval (the counting/tapping period) are unlikely to be recalled successfully in the subsequent cued-

recall test. To assess effects of forgetting that occurred after short-term repetition, we conducted an analysis of conditional delayed recall. This assessment of conditional recall performance allowed us to assess how much of the sentences were forgotten between the immediate and delayed recall tests, controlling for performance at immediate recall. Conditional recall was calculated simply as follows:

$$\text{Conditional recall} = \# \text{words in delayed condensed transcription} / \# \text{words in immediate condensed transcription}.$$

For the rare trials in which more words were recalled correctly in delayed than in immediate recall, we set this value to a maximum of 1.

Accuracy scores were averaged within subjects by condition and were subjected to a subject-wise repeated measures analysis of variance (ANOVA; F_1), with Sentence Type and Distraction Condition as within-subjects factors. Accuracy scores were also averaged across subjects for each experimental sentence and subjected to an item-wise ANOVA (F_2 ; Clark, 1973), with Distraction Condition as a repeated factor and Sentence Type as a between-items factor.

Gist To account for successful recall of information from the sentences that may have been missed by verbatim scoring, we also counted the number of “gist words” for which credit could be given. Acceptable gist words included synonyms or close semantic substitutions, morphological substitutions (e.g., verb tense changes), direct determiner substitutions, semantically close prepositional changes, and order changes, including active–passive voice alternations. Gist accuracy scores were analyzed statistically in the same manner as the verbatim scores.

Analysis of error types in recall

In addition to examining verbal responses for overall accuracy, we also measured the frequency of occurrence of the various kinds of errors. We defined six categories of errors: major order changes (correct words in the wrong order), unrelated additions, semantic substitutions, grammatical substitutions, phonological substitutions, and open-class omissions (open-class being defined as all words except closed-class words, which included prepositions, conjunctions, determiners, auxiliary verbs, and pronouns). See Appendix C for more details. The raw number of errors made by each participant in each category was submitted to a repeated measures ANOVA (F_1), with Sentence Type and Distraction Condition as within-subjects factors.

Assessment of interrater reliability

The majority of transcripts were analyzed by one rater for Experiment 1, and by two different raters for Experiment 2. Because some subjectivity was involved, particularly for the

assignment of gist points, we assessed interrater reliability for both the verbatim and gist scoring procedures. Three raters were trained in the scoring procedure, and each independently scored transcripts for three participants chosen randomly from Experiment 2. To assess reliability, we computed the intraclass correlation coefficient (ICC) by comparing the three raters' scores for each individual sentence in the short-term repetition and long-term recall tasks, pooling sentences across the three participants assessed, for a total of 300 sentences rated at both short-term and long-term recall. The specific algorithm used was "absolute two-way single measures ICC" (Hallgren, 2012). The ICC values were as follows: for short-term repetition, verbatim = .992, gist = .976; for raw long-term cued recall, verbatim = .980, gist = .960; and for conditional long-term cued recall, verbatim = .917, gist = .905. ICC values above .75 are generally considered "excellent" (Cicchetti, 1994).

Results

Experiment 1: short-term repetition and long-term cued recall with arithmetic distraction task

Short-term repetition: accuracy Both sentence type and distraction condition influenced participants' ability to repeat sentences following a filled delay. Recall performance across the four conditions is plotted in Fig. 2a (verbatim) and 2b (gist). The patterns of results were essentially identical using both verbatim and gist measures of recall for both the subject-wise and item-wise analyses. We found a main effect of sentence type [verbatim: $F_1(1, 19) = 35.33, p < .001, \eta_p^2 = .65$; $F_2(1, 98) = 13.01, p < .001, \eta_p^2 = .12$; gist: $F_1(1, 19) = 40.88, p < .001, \eta_p^2 = .68$; $F_2(1, 98) = 15.33, p < .001, \eta_p^2 = .14$], with participants achieving higher accuracy for concrete than for abstract sentences. We also observed a main effect of distraction condition [verbatim: $F_1(1, 19) = 71.09, p < .001, \eta_p^2 =$

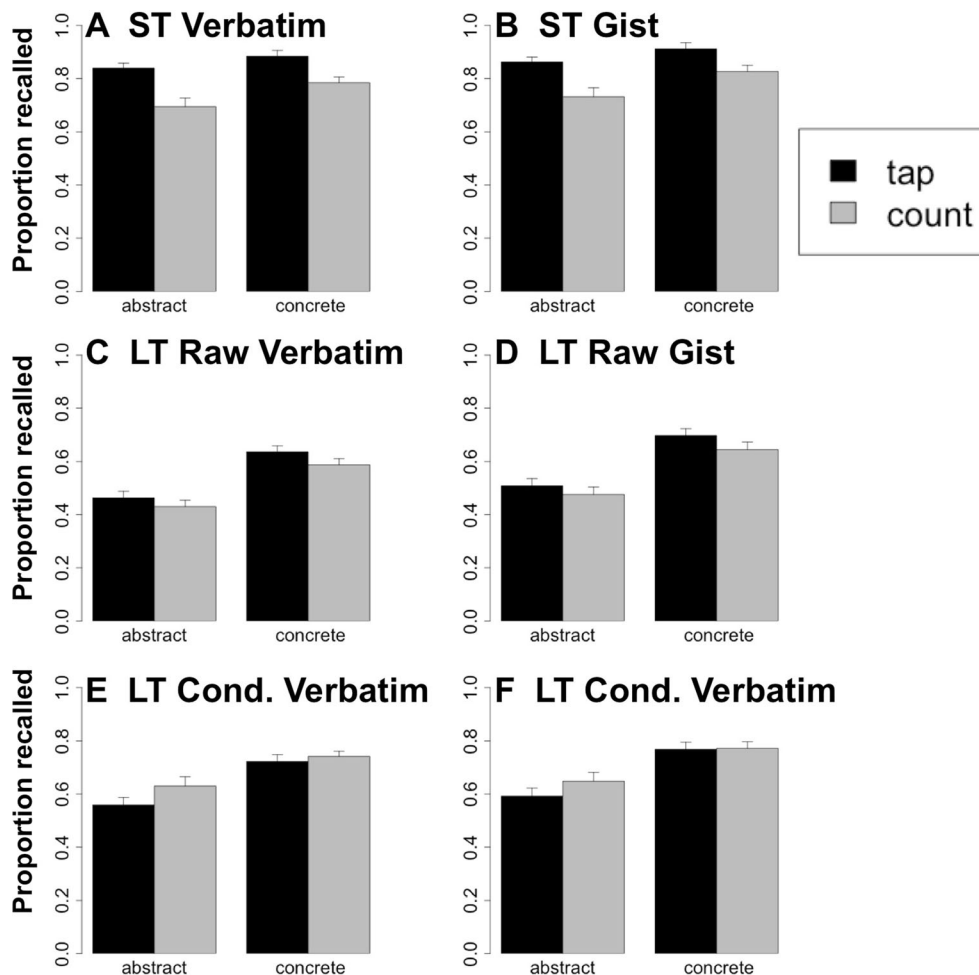


Fig. 2 Experiment 1: Recall accuracy. The proportions of words recalled accurately (# words recalled / # words in target sentence) in Task 1 in all four conditions are shown, for both the verbatim and gist scoring schemes. Error bars indicate one-sided 95 % confidence intervals

adjusted for repeated measures using the method of Morey (2008). (A–B) Recall accuracy in short-term (ST) repetition (Task 1). (C–D) Raw accuracy in long-term (LT) cued recall (Task 2). (E–F) Conditional (Cond.) recall in Task 2

.79; $F_2(1, 98) = 202.77, p < .001, \eta_p^2 = .67$; gist: $F_1(1, 19) = 52.57, p < .001, \eta_p^2 = .73$; $F_2(1, 98) = 160.52, p < .001, \eta_p^2 = .62$], with lower performance following counting than following finger tapping. Notably, the size of the effect for distraction condition was larger than that for sentence type; that is, in the counting condition, AS had a larger impact on performance than did sentence abstractness. Finally, a significant interaction was apparent between the two factors [verbatim: $F_1(1, 19) = 7.59, p = .012, \eta_p^2 = .29$; $F_2(1, 98) = 6.64, p = .011, \eta_p^2 = .06$; gist: $F_1(1, 19) = 6.40, p = .020, \eta_p^2 = .25$; $F_2(1, 98) = 7.17, p = .008, \eta_p^2 = .07$]. The form of the interaction was a stronger effect of distraction condition for abstract than for concrete sentences; that is, the AS caused by counting backward disrupted participants' ability to recall all of the sentences, but more strongly for abstract ones. Recall of concrete sentences was more resilient to the detrimental effects of AS, suggesting greater support from short-term maintenance of semantic information.

Long-term cued recall: raw accuracy Raw accuracy for the long-term cued-recall task is plotted in Fig. 2c (verbatim) and 2d (gist). We observed a main effect of sentence type [verbatim: $F_1(1, 19) = 100.13, p < .001, \eta_p^2 = .84$; $F_2(1, 98) = 57.38, p < .001, \eta_p^2 = .37$; gist: $F_1(1, 19) = 97.54, p < .001, \eta_p^2 = .84$; $F_2(1, 98) = 53.63, p < .001, \eta_p^2 = .35$], with participants recalling more words from concrete than from abstract sentences. There was also a main effect of distraction condition [verbatim: $F_1(1, 19) = 42.03, p < .001, \eta_p^2 = .69$; $F_2(1, 98) = 15.67, p < .001, \eta_p^2 = .14$; gist: $F_1(1, 19) = 29.19, p < .001, \eta_p^2 = .14$; $F_2(1, 98) = 15.20, p < .001, \eta_p^2 = .13$], with fewer words being remembered from sentences previously recalled following counting. No significant interaction emerged between the two factors [verbatim: $F_1(1, 19) = 1.04, p = .320, \eta_p^2 = .05$; $F_2(1, 98) = 0.71, p = .40, \eta_p^2 = .01$; gist: $F_1(1, 19) = 0.86, p = .36, \eta_p^2 = .04$; $F_2(1, 98) = 0.86, p = .35, \eta_p^2 = .01$].

Long-term cued recall: conditional accuracy Conditional accuracy for the long-term cued-recall task is plotted in Fig. 2e (verbatim) and 2f (gist). As in the immediate repetition task, we found a main effect of sentence type [verbatim: $F_1(1, 19) = 76.68, p < .001, \eta_p^2 = .80$; $F_2(1, 98) = 36.36, p < .001, \eta_p^2 = .27$; gist: $F_1(1, 19) = 82.80, p < .001, \eta_p^2 = .81$; $F_2(1, 98) = 39.18, p < .001, \eta_p^2 = .29$], with better recall for concrete sentences regardless of distraction condition. Not only were concrete sentences recalled better in short-term repetition, they were more resistant to being forgotten between immediate and long-term recall.

We also observed a main effect of distraction condition [verbatim: $F_1(1, 19) = 10.88, p = .003, \eta_p^2 = .36$; $F_2(1, 98) = 12.64, p < .001, \eta_p^2 = .11$; gist: $F_1(1, 19) = 4.51, p = .047, \eta_p^2 = .19$; $F_2(1, 98) = 5.23, p = .024, \eta_p^2 = .05$]. Interestingly, this effect was in the opposite direction of the effect upon immediate recall. Here, sentences that were previously

recalled under conditions of AS were relatively more preserved upon long-term recall; that is, they were forgotten less. This effect was stronger for the abstract sentences, as reflected by a significant interaction effect [verbatim: $F_1(1, 19) = 8.14, p = .010, \eta_p^2 = .30$; $F_2(1, 98) = 4.27, p = .041, \eta_p^2 = .04$; gist: $F_1(1, 19) = 6.17, p = .022, \eta_p^2 = .25$; $F_2(1, 98) = 4.54, p = .036, \eta_p^2 = .04$]. Considering the two sentence types alone, the effect of distraction condition was significant within abstract sentences [subject-wise paired t test: verbatim, $t(19) = 3.42, p = .003$; gist, $t(19) = 2.72, p = .014$], but not within concrete sentences [verbatim, $t(19) = 1.90, p = .073$; gist, $t(19) = 0.29, p = .77$]. These results indicate that AS prior to short-term repetition resulted in relatively better preservation of sentence content between short-term and long-term recall for abstract sentences, but had a lesser impact on subsequent recall for concrete sentences.

Short-term repetition: error type analysis The task conditions induced differential rates of errors within subjects, and the effects were different in the short-term and long-term recall tasks. For each participant, we counted the total number of errors in each category within each condition and subjected the totals to 2×2 repeated measures ANOVAs (subject-wise, not item-wise). Both major order errors and phonological errors were extremely rare and were not affected by any experimental factors, so they will not be discussed further. Figure 3 shows error rates across conditions in short-term repetition and long-term recall for three of the other error categories. Figure 3a shows *open-class omissions* during immediate recall. Omissions were by far the most common form of error, being words that were neither recalled successfully nor substituted. We found a main effect of sentence type [$F_1(1, 19) = 34.55, p < .001, \eta_p^2 = .65$] in which concrete sentences had fewer omissions, contributing to their higher overall accuracy. There was also a main effect of distraction condition [$F_1(1, 19) = 51.92, p < .001, \eta_p^2 = .73$], since sentences recalled after counting had more omissions. No interaction was present [$F_1(1, 19) = 2.02, p = .172, \eta_p^2 = .10$].

The same pattern of error rates in short-term repetition (both main effects significant and no interaction) was also observed for *unrelated additions* [Fig. 3b; sentence type, $F_1(1, 19) = 91.63, p < .001, \eta_p^2 = .83$; distraction condition, $F_1(1, 19) = 27.56, p < .001, \eta_p^2 = .59$; interaction, $F_1(1, 19) = 0.13, p = .721, \eta_p^2 = .01$]. For *semantic substitutions* (Fig. 3c), no main effect of sentence type emerged [$F_1(1, 19) = 1.55, p = .228, \eta_p^2 = .08$], but we did find an effect of distraction condition [$F_1(1, 19) = 31.62, p < .001, \eta_p^2 = .62$] and no interaction [$F_1(1, 19) = 0.60, p = .428, \eta_p^2 = .03$]. For *grammatical substitutions* (data not shown), the pattern of a significant main effect for distraction condition only was also present [sentence type, $F_1(1, 19) = 1.08, p = .312, \eta_p^2 = .05$; distraction condition, $F_1(1, 19) = 38.93, p < .001, \eta_p^2 = .67$; interaction, $F_1(1, 19) = 0.52, p = .478, \eta_p^2 = .03$].

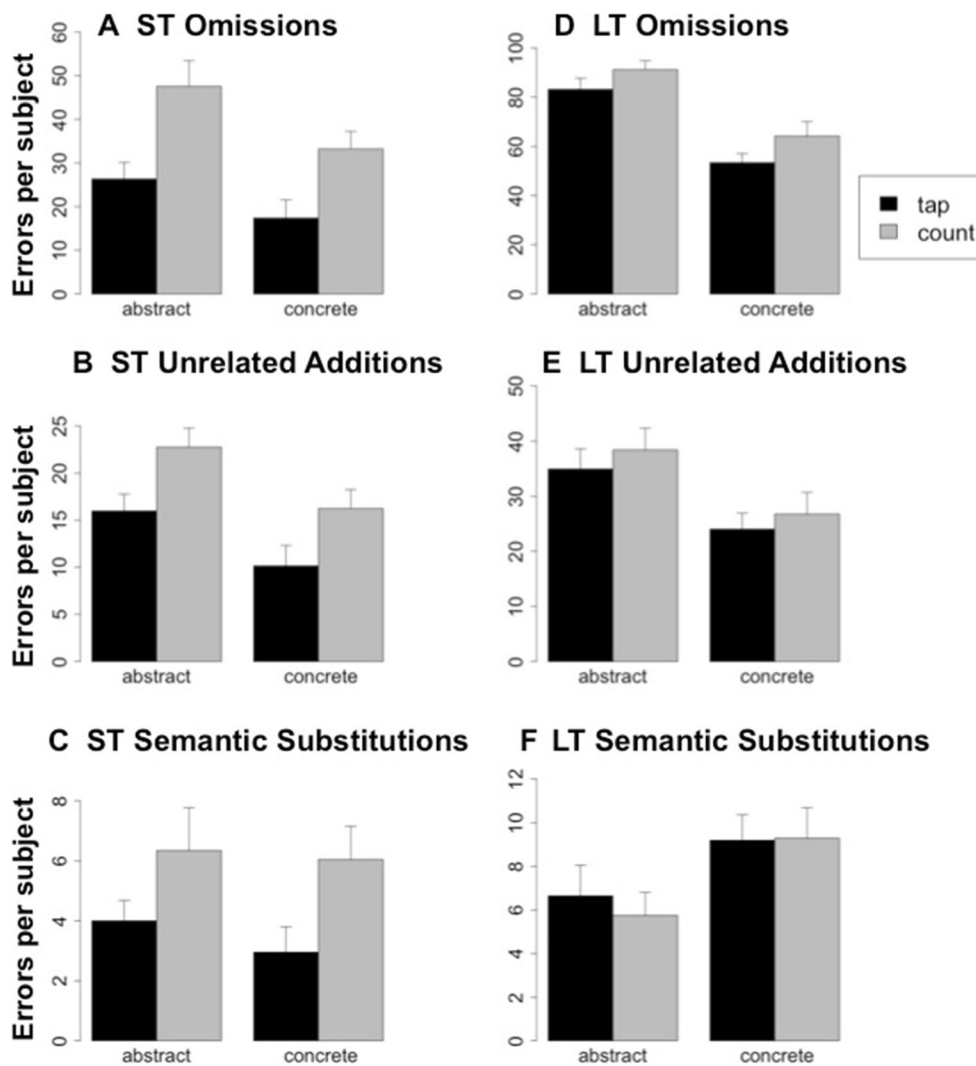


Fig. 3 Experiment 1: Occurrence of error types. These panels show the average numbers of errors of distinct types made by participants in the two tasks (total errors / # participants) for each condition. (A) Open-class omissions in short-term repetition. (B) Unrelated additions in short-term

repetition. (C) Semantic substitutions in short-term repetition. (D) Open-class omissions in long-term cued recall. (E) Unrelated additions in long-term cued recall. (F) Semantic substitutions in long-term cued recall

In summary, the pattern of errors in short-term repetition for most categories resembled that of accuracy in general. The manipulation of distraction condition had a stronger effect, with AS increasing the error count in four categories, while sentence type had a more modest effect, in that abstract sentences induced more errors in only two categories. However, a more varied pattern of error occurrence was present in the long-term cued-recall test.

Long-term cued recall: error type analysis In long-term cued recall, sentence type was a stronger modulator of error rates than was distraction condition, and the differences were not all in the same direction as they were in the immediate task. For *open-class omissions* (Fig. 3d), there were main effects of both sentence type [$F_1(1, 19) = 108.13, p < .001, \eta_p^2 = .85$] and distraction condition [$F_1(1, 19) = 32.96, p < .001, \eta_p^2 = .63$], but no interaction [$F_1(1, 19) = 0.53, p = .474, \eta_p^2 =$

.03]. For *unrelated additions* (Fig. 3e), we found a main effect of sentence type [$F_1(1, 19) = 40.32, p < .001, \eta_p^2 = .68$] and a marginal effect of distraction condition [$F_1(1, 19) = 4.25, p = .053, \eta_p^2 = .18$], with no interaction [$F_1(1, 19) = 0.03, p = .865, \eta_p^2 = .00$]. For both of these error types, errors were more frequent for abstract sentences and for sentences that had previously been recalled following AS. These effects resembled those seen for short-term repetition.

In contrast, *semantic substitutions* (Fig. 3f) had an opposite pattern. These errors were *more* common for concrete sentences [sentence type, $F_1(1, 19) = 26.20, p < .001, \eta_p^2 = .58$; distraction condition, $F_1(1, 19) = 0.53, p = .478, \eta_p^2 = .03$; interaction, $F_1(1, 19) = 0.57, p = .461, \eta_p^2 = .03$]. This pattern reflects the tendency for participants to maintain a gist meaning for concrete sentences more easily, which led to them generating words similar to the intended ones instead of omitting the target words altogether. This tendency was also

reflected in the accuracy data, since greater accuracy was consistently seen for concrete sentences, especially in the gist criteria that gave credit for semantic substitutions. For *grammatical substitutions* (not shown), no significant effects were found [sentence type, $F_1(1, 19) = 0.95, p = .343, \eta_p^2 = .05$; distraction condition, $F_1(1, 19) = 3.34, p = .083, \eta_p^2 = .15$; interaction, $F_1(1, 19) = 0.02, p = .884, \eta_p^2 = .00$].

Experiment 2: short-term repetition and long-term cued recall with simple AS

Short-term repetition: accuracy Recall performance across the four conditions is plotted in Fig. 4a (verbatim) and 4b (gist). Statistically, a main effect emerged of sentence type [verbatim: $F_1(1, 14) = 18.36, p < .001, \eta_p^2 = .57$; $F_2(1, 98) = 6.38, p = .013, \eta_p^2 = .06$; gist: $F_1(1, 14) = 16.47, p < .001, \eta_p^2 = .54$; $F_2(1, 98) = 7.37, p = .008, \eta_p^2 = .07$] and a main effect of distraction condition [verbatim: $F_1(1, 14) = 97.68, p < .001, \eta_p^2 = .87$; $F_2(1, 98) = 137.00, p = .011, \eta_p^2 = .58$; gist: $F_1(1, 14) = 68.59, p < .001, \eta_p^2 = .83$; $F_2(1, 98) = 121.00, p < .001, \eta_p^2 = .55$], but no interaction between the two factors

[verbatim: $F_1(1, 14) = 0.17, p = .687, \eta_p^2 = .01$; $F_2(1, 98) = 0.09, p = .760, \eta_p^2 = .00$; gist: $F_1(1, 14) = 1.21, p = .291, \eta_p^2 = .08$; $F_2(1, 98) = 0.92, p = .340, \eta_p^2 = .01$].

These results indicated that, as in Experiment 1, short-term repetition performance was better for concrete than for abstract sentences, and for sentences repeated after finger tapping rather than after AS. However, unlike in Experiment 1, the two factors did not interact: The detrimental effect of AS was approximately the same for both abstract and concrete sentences, whereas in Experiment 1, abstract sentences had been more vulnerable to disruption by AS than were concrete sentences.

Long-term cued recall: raw accuracy Raw accuracy for the long-term cued-recall task is plotted in Fig. 4c (verbatim) and 4d (gist). There was a main effect of sentence type [verbatim: $F_1(1, 14) = 44.33, p < .001, \eta_p^2 = .76$; $F_2(1, 98) = 41.02, p < .001, \eta_p^2 = .30$; gist: $F_1(1, 14) = 45.54, p < .001, \eta_p^2 = .76$; $F_2(1, 98) = 39.50, p < .001, \eta_p^2 = .29$], with concrete sentences being recalled more successfully than abstract ones. The effect of distraction condition was only marginal, reaching the $p <$

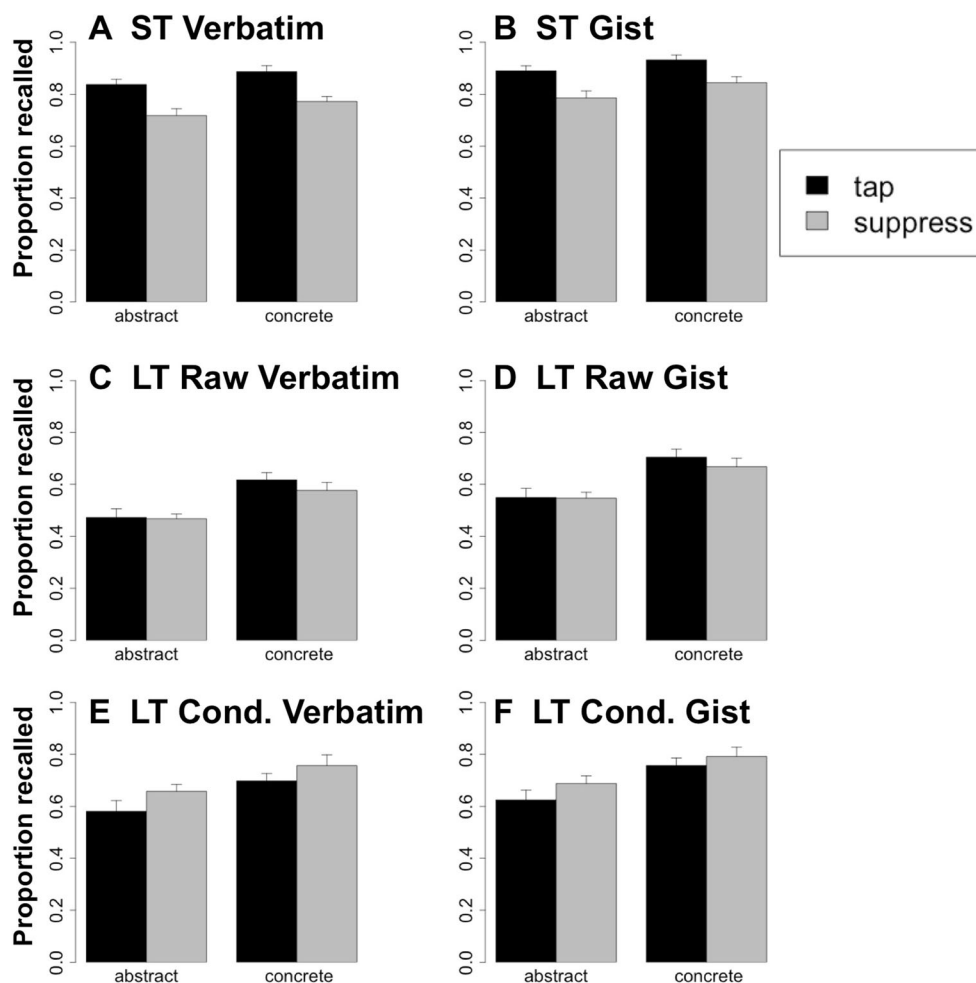


Fig. 4 Experiment 2: Recall accuracy. The interpretation of the panels is identical to that in Fig. 2

.05 criterion for verbatim but not for gist scoring [verbatim: $F_1(1, 14) = 5.15, p = .040, \eta_p^2 = .27$; $F_2(1, 98) = 4.35, p = .040, \eta_p^2 = .04$; gist: $F_1(1, 14) = 3.21, p = .095, \eta_p^2 = .19$; $F_2(1, 98) = 2.92, p = .091, \eta_p^2 = .03$], reflecting a slight decrement for sentences that were initially recalled following AS. Similarly, the interaction effect was marginal [verbatim: $F_1(1, 14) = 5.36, p = .036, \eta_p^2 = .28$; $F_2(1, 98) = 2.62, p = .109, \eta_p^2 = .03$; gist: $F_1(1, 14) = 5.36, p = .080, \eta_p^2 = .20$; $F_2(1, 98) = 2.00, p = .160, \eta_p^2 = .02$].

Long-term cued recall: conditional accuracy Conditional accuracy for the long-term cued-recall task is plotted in Fig. 4e (verbatim) and 4f (gist). We found a main effect of sentence type [verbatim: $F_1(1, 14) = 20.37, p < .001, \eta_p^2 = .59$; $F_2(1, 98) = 33.42, p < .001, \eta_p^2 = .25$; gist: $F_1(1, 14) = 26.56, p < .001, \eta_p^2 = .65$; $F_2(1, 98) = 36.16, p < .001, \eta_p^2 = .27$] and a main effect of distraction condition [verbatim: $F_1(1, 14) = 38.06, p < .001, \eta_p^2 = .73$; $F_2(1, 98) = 21.42, p < .001, \eta_p^2 = .18$; gist: $F_1(1, 14) = 45.77, p < .001, \eta_p^2 = .77$; $F_2(1, 98) = 10.89, p = .001, \eta_p^2 = .10$], but no interaction [verbatim: $F_1(1, 14) = 0.82, p = .381, \eta_p^2 = .06$; $F_2(1, 98) = 2.34, p = .129, \eta_p^2 =$

.02; gist: $F_1(1, 14) = 1.36, p < .264, \eta_p^2 = .09$; $F_2(1, 98) = 3.56, p = .062, \eta_p^2 = .04$].

These results indicated that, as in Experiment 1, concrete sentences were better preserved (i.e., less forgotten) than abstract sentences between short-term repetition and long-term cued recall for the same sentences. Similarly, sentences repeated after AS in the short-term repetition task were better preserved than those repeated after finger tapping, when tested in long-term cued recall. Unlike in Experiment 1, the two factors did not interact: The better preservation of sentences initially repeated after AS was equivalent in size for both concrete and abstract sentences.

Short-term repetition: error type analysis Figure 5a–c show error rates across conditions in short-term repetition for three of the error categories. The overall pattern was identical to that in Experiment 1. We observed very few major order or phonological errors, and no significant effects of the experimental factors on these types. The remaining four error types were all more common following abstract sentences and following AS, and no interactions were present except for

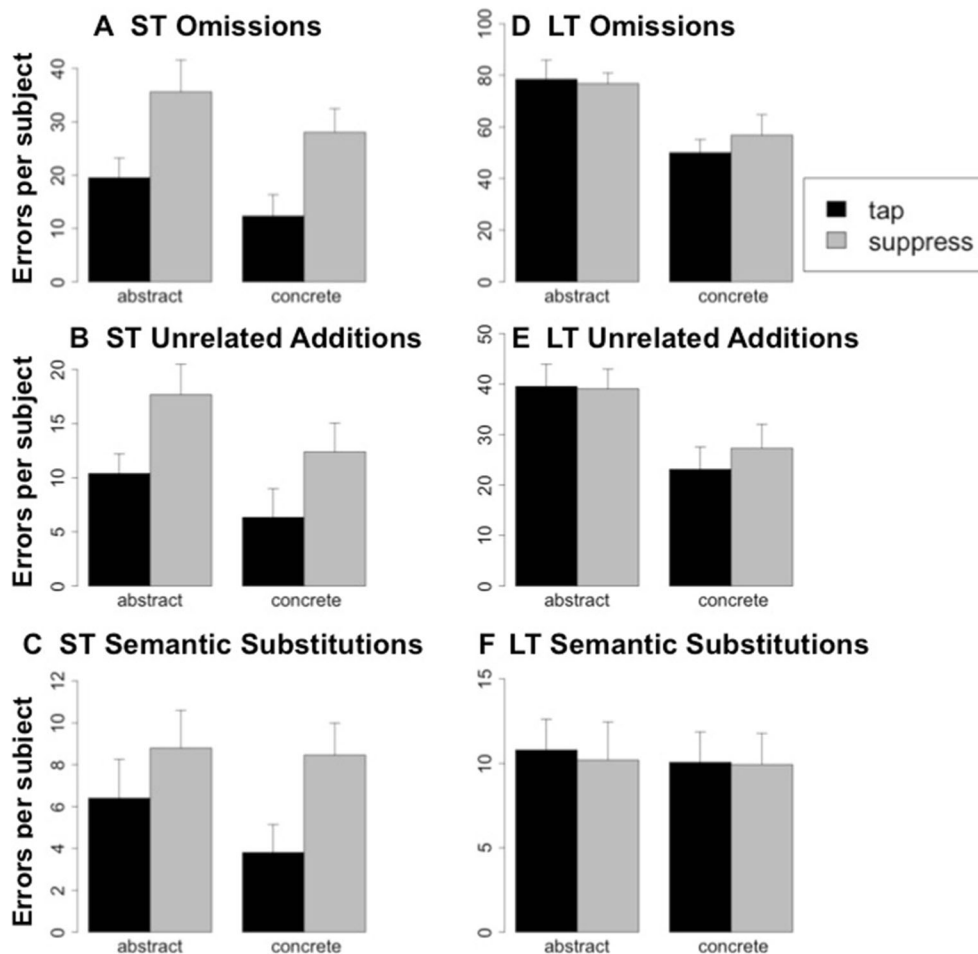


Fig. 5 Experiment 2: Occurrence of error types. The interpretation of the panels is identical to that in Fig. 3

grammatical substitutions. The ANOVA results were as follows: *Open-class omissions* (Fig. 5a): sentence type, $F_1(1, 14) = 13.19, p = .003, \eta_p^2 = .49$; distraction condition, $F_1(1, 14) = 32.43, p < .001, \eta_p^2 = .70$; interaction, $F_1(1, 14) = 0.02, p = .881, \eta_p^2 = .00$; *unrelated additions* (Fig. 5b): sentence type, $F_1(1, 14) = 29.73, p < .001, \eta_p^2 = .68$; distraction condition, $F_1(1, 14) = 18.13, p < .001, \eta_p^2 = .56$; interaction, $F_1(1, 14) = 0.38, p = .549, \eta_p^2 = .03$; *semantic substitutions* (Fig. 5c): sentence type, $F_1(1, 14) = 5.06, p = .041, \eta_p^2 = .27$; distraction condition, $F_1(1, 14) = 16.46, p < .001, \eta_p^2 = .54$; interaction, $F_1(1, 14) = 2.23, p < .158, \eta_p^2 = .14$; *grammatical substitutions* (data not shown): sentence type, $F_1(1, 14) = 13.31, p = .003, \eta_p^2 = .49$; distraction condition, $F_1(1, 14) = 14.03, p = .002, \eta_p^2 = .50$; interaction, $F_1(1, 14) = 5.47, p = .035, \eta_p^2 = .28$.

Long-term cued recall: error type analysis Figure 5d–f shows error rates in long-term cued recall for Experiment 2. Again, no significant effects emerged for major order and phonological errors (data not shown). Sentence type exerted an effect, with more errors in three categories (open-class omissions, unrelated additions, and grammatical substitutions) for abstract sentences. No significant effects of distraction condition were present, however; all error categories were equally frequent for sentences previously repeated following both finger tapping and AS. As in Experiment 1, semantic substitutions patterned differently from the other common error types: In this case, there were no significant effects of either factor on the occurrence of semantic substitutions (whereas in Exp. 1, semantic substitutions were *more* common for concrete sentences, and other error types were less common for concrete sentences).

The ANOVA results were as follows: *Open-class omissions* (Fig. 5d): sentence type, $F_1(1, 14) = 39.52, p < .001, \eta_p^2 = .74$; distraction condition, $F_1(1, 14) = 0.82, p = .381, \eta_p^2 = .06$; interaction, $F_1(1, 14) = 4.95, p = .043, \eta_p^2 = .26$; *unrelated additions* (Fig. 5e): sentence type, $F_1(1, 14) = 29.47, p < .001, \eta_p^2 = .68$; distraction condition, $F_1(1, 14) = 3.41, p = .086, \eta_p^2 = .20$; interaction, $F_1(1, 14) = 1.14, p = .304, \eta_p^2 = .08$; *semantic substitutions* (Fig. 5f): sentence type, $F_1(1, 14) = 0.24, p = .634, \eta_p^2 = .02$; distraction condition, $F_1(1, 14) = 0.25, p = .625, \eta_p^2 = .02$; interaction, $F_1(1, 14) = 0.06, p < .806, \eta_p^2 = .00$; *grammatical substitutions* (data not shown): sentence type, $F_1(1, 14) = 53.37, p < .001, \eta_p^2 = .79$; distraction condition, $F_1(1, 14) = 0.93, p = .352, \eta_p^2 = .06$; interaction, $F_1(1, 14) = 0.93, p = .352, \eta_p^2 = .06$.

Discussion

This study introduced a novel paradigm for assessing sentence recall at both short-term and long-term stages. We selectively interfered with subvocal rehearsal on some trials by using AS

tasks. Experiment 1 was based on a relatively demanding backward-counting task, whereas Experiment 2 was based on a simpler nonword articulation task. The predictions for short-term repetition were straightforward and were confirmed for both experimental factors. Concrete sentences were repeated more accurately than abstract sentences, reflecting the greater support available from the semantic mechanisms that complement pSTM. The AS tasks reduced accuracy for short-term repetition, as compared to finger tapping, during the short-term delay period. This reduction in accuracy is attributable to interference with pSTM, which ordinarily supports relatively high performance in short-term sentence repetition in the absence of distraction. Finally, AS tasks caused a larger performance decrement for abstract than for concrete sentences in Experiment 1, reflecting the lesser semantic support available for these sentences and their increased reliance on pSTM to support accurate repetition. These effects illustrate that semantic mechanisms play a role in short-term repetition that complements pSTM, especially following a brief delay.

Of greater interest for the present study are the effects of the experimental manipulations on subsequent cued recall of the same sentences in Task 2. This experiment tested whether subvocal rehearsal during a short-term maintenance period promotes the retention of sentence content in LTM when it is tested later, relative to conditions in which rehearsal is prevented and in which semantic mechanisms may be selectively engaged in Task 1 to support repetition in the absence of rehearsal. According to the rehearsal advantage account, sentences initially repeated after AS could be subject to greater rates of forgetting between the two tasks. Alternatively, according to a levels-of-processing framework, the engagement of semantic mechanisms under conditions of AS could result in more effective encoding into LTM, thus protecting those sentences from forgetting. Critically, the two positions diverge in their predictions for conditional recall, or the number of words recalled in Task 2 relative to the number recalled for the same sentence in Task 1. Because we did not expect words forgotten in Task 1 to be recalled in Task 2, we focused on conditional recall as a measure of forgetting after Task 1.

The results clearly support the levels-of-processing prediction. In both experiments, conditional recall was better for sentences that were initially recalled following a challenging distractor task that interfered with subvocal rehearsal. That is, AS before short-term repetition led to sentences being forgotten less between their first recall attempt (short-term repetition, Task 1) and their second recall attempt (long-term cued recall, Task 2). Additionally, concrete sentences were remembered better than abstract sentences at both stages: They were recalled more accurately in Task 1, and they were forgotten less in Task 2.

These results show a dissociation between the two experimental factors. Whereas sentence concreteness had beneficial

effects on recall performance in both short-term repetition (Task 1) and conditional long-term cued recall (Task 2), the effects of distraction condition were reversed between the two tasks: AS resulted in sentences being forgotten more in Task 1, but being forgotten less between the two tasks. Despite this dissociation, all of these effects are plausibly attributable to the engagement of semantic mechanisms in short-term repetition that complement pSTM. Concrete sentences can more easily be maintained in memory through semantic resources, including visual imagery and schema construction. Although pSTM supports verbatim maintenance of sentences under ideal conditions (without distraction), retrieval of sentences' meanings from cSTM or LTM can support the regeneration, or "redintegration," of the phonological form, resulting in relatively good short-term recall following distraction.

In the present study, we hypothesized that the engagement of semantic mechanisms in the retrieval of sentence content would be greater when pSTM was blocked by AS, resulting in a benefit for LTM. The reversal of the distraction effect between Tasks 1 and 2 supports this hypothesis. Although blocking pSTM reduces performance in short-term repetition, the redintegration of the sentence following the distraction results in deeper processing of the sentence's meaning, such that it is less likely to have been forgotten when it is cued in Task 2. This result may be somewhat surprising, since rehearsal of a sentence's form in pSTM would ordinarily be expected to have a beneficial effect on subsequent memory relative to the absence of rehearsal. However, the present results suggest that phonological rehearsal, despite its effectiveness for short-term repetition, may be a "shallower" form of maintenance that does not contribute much to LTM encoding. Semantic elaboration results in better LTM encoding, and semantic elaboration is enhanced when pSTM is blocked but the task demands nonetheless require the sentence's retrieval for short-term repetition.

One potential concern about these findings is that a long-term advantage for the AS condition was only found for conditional delayed recall, and not for raw delayed recall. We did not expect to find an effect for raw recall, because words that are already forgotten in Task 1 are unlikely to be recalled in Task 2. However, one might argue that the beneficial effect of AS on Task 2 is an artifact of the conditional-recall scoring procedure. Since accuracy on Task 2 was compared with accuracy for the same sentence in Task 1, any manipulation that decreased performance in Task 1 (such as AS) might be expected to increase the conditional recall score on Task 2 if it actually had no effect on the degree to which the sentence was encoded into LTM. However, the present results are unlikely to be attributable to such an effect alone, because the reversal was only seen for the effect of distraction condition. Sentence abstractness resulted in decreased accuracy in short-term repetition and *also* decreased accuracy in conditional recall. That is, abstract sentences were forgotten more than concrete ones

during the brief delay period before their first recall attempt, and then *further* forgotten before their second recall attempt. This is the opposite of the pattern seen for AS, suggesting that decreased conditional recall is not an inevitable consequence of increased short-term recall. Rather, both effects are more easily explained by the increased engagement of semantic resources for short-term repetition when pSTM is blocked. This finding indicates that high performance in immediate recall, when driven by phonological rehearsal, is not necessarily predictive of good encoding into LTM, which is consistent with the popular notion that "rote" rehearsal may not be the most effective technique for the memorization of verbal information.

Besides the findings about accuracy, the patterns of recall errors observed in these experiments support the idea that semantic engagement in short-term repetition supports subsequent accuracy in long-term cued recall. In short-term recall, all error types (except those that were completely unaffected by the experimental manipulations) were more frequent in conditions that reduced overall accuracy: counting versus tapping and abstract versus concrete sentences. In long-term recall, omissions and unrelated additions were more frequent for abstract sentences (patterning with overall accuracy), but semantic substitutions showed a different pattern. In Experiment 1, the effect of sentence type was actually reversed for semantic substitutions relative to other error types: They were significantly more frequent for concrete than for abstract sentences. In Experiment 2, the effect did not quite reverse as compared to other error types, but it was neutralized: Semantic substitution errors were equally frequent for both abstract and concrete sentences.

Because concrete sentences are thought to be more amenable to semantic encoding (and therefore less dependent on phonological rehearsal to support immediate recall), redintegration of a sentence's phonological form on the basis of its meaning is likely to result in semantic substitutions. At the delayed stage, the phonological form is redintegrated solely from the retained meaning, and thus semantic effects dominate: an increase in semantic substitutions for better-remembered sentences, and a decrease in other error types. In short-term repetition, in contrast, all error types are decreased for concrete sentences. This suggests that the phonological trace still contributes to recall performance in short-term repetition, even in the face of AS, although semantic encoding does contribute.

Differences related to the type of AS

In the two experiments, we used different tasks for AS. In Experiment 1, we used a fairly demanding cognitive task, counting backward by threes. We chose this task because we not only wanted to block subvocal rehearsal; we also wanted to prevent participants from covertly refreshing the

phonological form of the sentences between utterances, since an insufficiently demanding distractor task might allow for the latter (Rose et al., 2014). However, this task might also have interfered with other cognitive resources that could contribute to memory encoding, including attention and executive processes. Thus, the beneficial effects of distraction on conditional long-term recall seen in Experiment 1 might be expected to be eliminated if a simpler AS task were used, because refreshing might suffice to keep the phonological form of sentences in active memory. Alternatively, the effects might be enhanced if the relative preservation of attention and executive functions makes a strong contribution to encoding in this case. In fact, the results of both experiments were very similar: In both experiments, AS reduced accuracy in short-term repetition but improved conditional accuracy in long-term cued recall. One notable difference between the two experiments was the interaction between imageability and distraction condition. Significant interactions were apparent for both tasks in Experiment 1: Abstract sentences suffered more than concrete sentences from AS on short-term repetition, but benefited more from it on delayed recall, as predicted. In Experiment 2, however, these interactions were not present; AS reduced accuracy equivalently for both kinds of sentences on short-term repetition, and equivalently improved accuracy for long-term cued recall. An additional difference between the two experiments was the relative frequencies of semantic substitutions for concrete versus abstract sentences in delayed recall, as we noted in the previous section.

The similar results seen in both experiments suggest that traditional AS is sufficient to disrupt pSTM and cause participants to rely more on semantic resources to support recall in short-term repetition. The additional cognitive load caused by verbal calculation in Experiment 1 did not appear to interfere with the engagement of those semantic resources. In fact, it seemed to enhance the effects, bringing about the expected interaction between imageability and distraction condition in Experiment 1, and reversing the frequency of semantic substitutions between the two tasks. These enhanced effects may be attributable to a more demanding distraction task blocking covert retrieval of a sentence's phonetic form, which participants may have been able to do occasionally between repetitions of the nonword "Babataka" in Experiment 2.

Implications for neuropsychology

The findings in this study suggest that rehearsal of verbal information in the phonological loop is not necessarily optimal for encoding information into LTM. Although unrestricted pSTM supports very high performance in short-term repetition, we have found that interfering with pSTM through AS tasks results in less forgetting of sentence content between the short-term repetition and long-term cued-recall tests.

The precise nature of the mechanisms that supplement pSTM for sentence repetition remains controversial. Although some behavioral data do suggest that there are dedicated mechanisms for the short-term maintenance of semantic information (e.g., Romani, McAlpine, & Martin, 2008; Shivde & Anderson, 2011), some theorists maintain that short-term maintenance of verbal information is attributable to the encoding and retrieval of information into LTM on a short time scale (see the introduction). The neuropsychological implication of such a view would be that both short- and long-term verbal memory depend on the same brain structures, most likely the medial temporal lobe (MTL). On the other hand, dedicated maintenance of semantic information is more likely to be related to brain activity in cortical areas previously linked to semantic processing (Binder, Desai, Graves, & Conant, 2009), especially the anterior temporal lobes (Patterson, Nestor, & Rogers, 2007).

Some evidence for a distinct semantic STM buffer has come from neuropsychological cases demonstrating a double dissociation between deficits in phonological and semantic STM (Belleville, Caza, & Peretz, 2003; N. Martin & Saffran, 1997; R. C. Martin & He, 2004; R. C. Martin, Shelton, & Yaffee, 1994). Although these findings support a strong dissociation between phonological and semantic resources for sentence repetition, they are compatible with a critical role for the MTL memory systems in supporting the use of semantic information in STM for sentence content. However, more recent neuropsychological evidence from MTL amnesic patients suggests that semantic support in verbal STM is also independent from MTL-mediated episodic memory function (Race, Palombo, Cadden, Burke, & Verfaellie, 2015; Rose, Olsen, Craik, & Rosenbaum, 2012). These findings suggest that neocortical regions distinct from those underlying pSTM may support specialized resources that are particularly important for semantic STM. Characterizing the relevant brain networks remains an important task for future studies.

Implications for neuroimaging

A few neuroimaging studies have explored the brain mechanisms supporting semantic STM. Shivde and Thompson-Schill (2004) specifically implicated the bilateral inferior frontal and left middle temporal gyri in semantic maintenance, whereas Hamilton, Martin, and Burton (2009) had partially overlapping findings implicating the left inferior and middle frontal gyri. These findings of cortical activation are consistent with the existence of specialized mechanisms for semantic maintenance, although they may also be attributable to increased semantic *processing* rather than to maintenance per se. Recently, Rose, Craik, and Buchsbaum (2015) found that short-term recall accuracy for single words was predicted by the level of activity in the left inferior frontal gyrus at encoding, in the left anterior temporal lobe during maintenance of the word in a distractor-filled delay period, and in the left hippocampus following distraction (i.e.,

during the recall phase itself). Critically, the recruitment of these areas that are commonly involved in semantic processing and episodic recall depended on the nature of the distraction; rehearsal-filled delays recruited a very different network of frontal, temporal, and parietal areas associated with the default-mode network. To investigate the question of whether semantic maintenance is driven by the engagement of LTM systems operating over short time scales, or by dedicated short-term mechanisms, it would be desirable to examine both short-term and long-term sentence recall using neuroimaging techniques.

The cued sentence recall paradigm developed here may also prove useful in neuroimaging studies of verbal memory, particularly regarding the distinctions between pSTM, cSTM, and LTM. Despite the hundreds of fMRI studies that have been conducted in the past two decades, a major debate persists between single-mechanism and dual-mechanism accounts of STM and LTM (Ranganath & Blumenfeld, 2005; Surprenant & Neath, 2008). Some studies have attempted to address the question by conducting both working memory and long-term encoding or retrieval tasks within the same participants, and have found both overlapping and distinct activations (Braver et al., 2001; Cabeza, Dolcos, Graham, & Nyberg, 2002).

One of the more powerful techniques for assessing neural correlates of the encoding of information into LTM is the subsequent memory technique. Stimuli are initially presented for encoding (either intentional or incidental) during neural data collection, and then either recognition or recall for the same stimuli is assessed afterward. Neural activity is then compared for stimuli that were subsequently remembered or forgotten. Since its initial use in event-related fMRI (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Wagner, Koutstaal, & Schacter, 1999), the paradigm has been used in dozens of fMRI experiments. A meta-analytic review (Kim, 2011) revealed that activation predicting subsequent memory for verbal stimuli tended to be highly left-lateralized, occurring most commonly in the medial temporal lobe, inferior temporal gyrus, and inferior frontal gyrus, all regions that are intimately linked with semantic processing. Negative effects (activity predicting subsequent forgetting) occurred in default-mode structures.

Despite the power of this technique to elucidate the underpinnings of LTM encoding, we know of no published studies that have used it to evaluate the relationships between verbal STM and LTM. This may be due to the lack of an appropriate task. The studies reviewed have tended to use single words or word pairs for encoding, and either recognition or free recall for assessing memory performance. Single words or word pairs are not very challenging as stimuli for STM, and combining them into a longer list makes it difficult to assess subsequent memory for a particular item occurring in a neuroimaging study of STM. We suggest that the sentence is a natural

unit for assessing both STM and LTM, and the cued-recall paradigm developed here could be ideal for neuroimaging investigations of the relationship between memory on these two time scales. Furthermore, we expect that such studies would not simply replicate prior subsequent memory studies, because they would allow researchers to manipulate effects of phonological rehearsal and AS. In the present study, we have seen a competitive interaction between phonological rehearsal and semantic processing, such that interfering with phonological rehearsal resulted in less forgetting of sentences at the delayed stage. Thus, one might predict that areas implementing phonological rehearsal would show a negative subsequent-memory effect, since they are more active during STM for sentences that are subsequently forgotten, reflecting their role in relatively shallow “rote” memory that does not involve deep processing of semantic content. Similarly, areas involved in semantic maintenance may show a positive subsequent-memory effect. Future experiments examining both STM and LTM for sentence stimuli may therefore be able to dissociate the distinct roles of specific brain regions in the phonological and semantic short-term memory of verbal information.

Author note This research was supported by a grant from the Heart and Stroke Foundation Canadian Partnership for Stroke Recovery, and by an Alzheimer’s Association New Investigator Research Grant to J.A.M.

Appendix A: Imageability rating survey

Raters of candidate sentences were given the following written instructions:

We would like you to rate the following sentences on “imageability.” This refers to how easily you can form a mental picture of the sentence’s meaning. For example, a sentence like this would typically be rated rather high in imageability:

“A little boy has 3 red marbles in his hand.”

Here, you can form a visual image in your mind of a little boy and “see” the three red marbles in his hand. With this image in mind, you might even be able to draw a picture and get someone else to guess something close to the original sentence by looking at your picture. In contrast, the following sentence would be considered low in imageability:

“For every rule, there is an exception.”

The meaning of this sentence is clear, but it is not as easy to visualize. It would probably be rather difficult to transmit the sentence’s meaning through any means other than words.

Please rate each sentence on a scale of 1 to 5, with 5 being *highly imageable* (red marbles) and 1 being *minimally imageable* (rules and exceptions).

Appendix B: Quantitative matching of sentences

The 100 sentences selected for the study were divided into two sets of 25 abstract, and two sets of 25 concrete sentences. The two sets were alternately assigned to the tapping condition and the counting condition in a counterbalanced fashion across participants. We furthermore attempted to match the different sets on a number of quantitative criteria that might affect recall performance. Although such matching is easily done on individual words using published norms, it is considerably more complicated for sentences. To minimize the likelihood of incidental differences between the sentence sets, we employed some readily available tools from the computational-linguistics literature to compute metrics of sentence complexity and predictability. Sentences from each condition were randomly assigned into sets of 25, and the degree of match was assessed. The random selection was repeated until satisfactory matching was obtained. For all measures described below, the two randomly constructed sets assigned to either tapping or counting (counterbalanced across participants) within each sentence category (abstract or concrete) were well matched ($p > .05$ on an unpaired t test). However, some systematic differences between abstract and concrete sentences are reflected in these measures. These quantitative differences are related to normal differences between abstract and concrete sentences in English, and attempting to design them out of the sentences would have sacrificed their ecological validity. Nonetheless, we discuss them in detail below and how they are unlikely to confound our results. Quantitative analysis of the sentences was implemented in scripts written in Python and R, making extensive use of the Natural Language ToolKit software (NLTK; Bird, Klein, & Loper, 2009; <http://nltk.org/>). Other tools used are described below. Measures computed included the following:

Length

The length of a sentence can be defined in various ways: words, syllables, or actual spoken duration. Because words for abstract concepts in English tend to be longer than concrete words, it is almost impossible to match a large number of abstract sentences with concrete sentences on both words and syllables. Behavioral evidence suggests that the capacity for phonological rehearsal is sensitive to the time it takes to covertly

articulate materials, corresponding to such variables as speech rate (Hulme et al., 1991) and word length (Baddeley, Thomson, & Buchanan, 1975; Service, 1998; but see Bireta, Neath, & Surprenant, 2006, for a review of the nuances of word length effects). To control for phonological STM demand independent of an individual's speech rate, we decided that it was most important to match sentences on length in syllables. Concrete and abstract sentences did not differ significantly in syllable length (means 20.20 and 20.54), but the concrete sentences tended to be one word longer, on average (13.10 vs. 12.12, two-sample t test $p < .01$). This was largely due to the increased presence of small function words in the concrete sentences. Many of the concrete sentences described motion events that frequently rely on phrasal verbs (e.g., “The ship *deviated from* its course during the night and *collided with* an iceberg”), but such constructions were less frequent in abstract sentences. This difference likely accounts for several of the other quantitative differences described below. Semiautomated syllable counting was facilitated by the CMU pronunciation dictionary (www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed 2012), interfaced via NLTK.

Word frequency

We quantified word frequency (WF) using the norms of Kučera and Francis (1967), in two different ways. We calculated an average *inclusive* WF, for which every word in the sentences was counted, and an *exclusive* WF, for which a set of common closed-class “stopwords” in English was excluded from the calculation. Both of these measures differed between concrete and abstract sentences, but in *opposite* directions: The inclusive WF was higher for concrete sentences (7,023 vs. 5,969, $p = .014$), but the exclusive WF was higher for abstract sentences (90 vs. 158, $p = .027$). Again, this reflects the fact that concrete sentences tend to contain more of the common function words, but the content words in the abstract sentences tend to be of higher frequency than the content words of concrete sentences. This may be attributable to the authors' selection of highly imageable objects that are relatively uncommon in daily life for the concrete sentences.

Syntactic complexity

More complex sentences have been shown to be recalled less accurately. Cheung and Kemper (1992) found that two metrics, in particular, best accounted for participants' difficulties in comprehending and repeating sentences: number of clauses and Yngve depth

(Yngve, 1960). The latter measure quantifies the extent to which the syntactic structure of a sentence contains left- rather than right-branching phrases. For detailed illustrations of how Yngve depth is quantified, see Yngve (1960), Cheung and Kemper (1992), and Sampson (1997). We quantified Yngve depth as both the mean depth over all words and the maximum depth in the sentence. For example, a sentence with an object-embedded relative clause, such as “The juice that the child spilled stained the rug,” is more left-branching than one with a subject-embedded relative clause, such as “The child spilled the juice that stained the rug” (Stromswold, Caplan, Alpert, & Rauch, 1996). Using our procedures, the first sentence was assigned the values [max depth: 3, mean depth: 1.67], while the second sentence was assigned [max depth: 2, mean depth: 1.11].

To calculate syntactic complexity, we submitted all sentences to automated phrase structure analysis using the Stanford parser (Klein & Manning, 2003). The number of clauses was defined as the number of nodes in each sentence’s parse tree assigned the tag of “S” (declarative clause) according to the annotation system of the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993). Concrete and abstract sentences did not differ significantly in numbers on this measure (1.68 vs. 1.80, $p = .48$). The total numbers of phrasal nodes in the parse trees also did not differ between conditions (10.0 vs. 10.8, $p = .50$). Both metrics of Yngve depth indicated a slight increase in syntactic complexity for concrete sentences, of marginal statistical significance (mean depth, 1.57 vs. 1.47, $p = .046$; max depth, 3.20 vs. 2.92, $p = .057$). This difference indicates that concrete sentences tended to have slightly more complex phrase structures, perhaps due to the increased presence of short, lexicalized phrases. However, syntactic complexity alone would predict poorer recall for concrete sentences, when in fact, as we found, the opposite was true. Therefore, syntactic complexity does not seem to confound the recall advantage for concrete sentences, although it may have reduced the size of the effect by acting in the opposite direction.

Predictability

Sentence recall spans are much longer than span for unrelated word lists, presumably due to the support given by semantic processing, which causes later words to be predictable from earlier words. Therefore, it was desirable to control for predictability between conditions. Predictability depends strongly on the meaning of a sentence as a whole (not on the sum of the individual words), and is thus extremely difficult to quantify, except in highly constrained cases, such as in the cloze

procedure (Taylor, 1953). However, an approximate measure can be derived from a statistical analysis of natural texts. In particular, n -gram models of language are widely used in computational linguistics to quantify the amount of information and redundancy in natural text. In essence, these models compute the conditional probability of each word, given n preceding words. We used the CMU–Cambridge Statistical Language Modeling Toolkit (Clarkson & Rosenfeld, 1997) to construct a 5-gram model of English, based on an eight-million-word sample of natural texts, comprising a subset of the American National Corpus (Ide & MacLeod, 2001) that excluded portions of that corpus derived from technical biomedical texts. We computed 5-grams (and smaller n -grams) from the 20,000 most common words of the sample. For each sentence, we computed the total entropy summed over words measured in bits, as well as the more common metric of perplexity (2^H , where H is the entropy). Concrete and abstract sentences did not differ in either of these measures (entropy, 9.32 vs. 9.26, $p = .87$; perplexity, 932 vs. 823, $p = .61$). This indicates that the two sentence types were roughly equivalent in predictability, as far as can be determined by statistical language modeling.

Appendix C: Scoring procedures

Condensed verbatim transcription

All verbal responses were first transcribed exactly as uttered by the participant. To calculate a verbatim accuracy score for each response, the transcription was first compared to the target sentence, and a “condensed” transcription consisting of only words that were judged to be accurately recalled verbatim was produced.

The following conditions had to be met for a word to be included:

1. The word had to be in the same tense as in the original sentence.
2. The word could not be pluralized or altered in any way (i.e., it had to be the exact same word as in the original sentence).
3. Words did not necessarily need to be in the correct order, but a word **MUST** be the exact same word present in the original sentence.
4. If the participant recalled a completely different clause or proposition but remembered a word present in the original sentence (such as “an,” “or,” etc.), this word was transcribed.
5. Words were not included if they were not part of the original sentence.

Examples:

Target Sentence	Subject's Recall Attempt	Condensed Verbatim Transcription
The snake handler wrapped his finger in a Band-Aid after getting a nasty bite.	The snake handler wrapped his wrist after a particularly nasty snake bite.	The snake handler wrapped his after a nasty bite.
Rainbows have colored the sky this week during the ongoing thunderstorms.	Rainbows colored the sky after a whole week of thunderstorms.	Rainbows colored the sky week thunderstorms.
The hostess ate all the leftover food after the party guests had all gone home.	The hostess ate all of the leftover food after the guests had left the party.	The hostess ate all the leftover food after the party guests had.

Assignment of gist points

Using the verbatim score (# words in condensed verbatim transcription) as a starting point, a gist score was computed by adding additional points for words that helped expressed the meaning of the target sentence.

Gist words that were given 1 point per word:

1. **Synonyms or semantic substitutions** were given 1 point if they retained the overall meaning of the sentence and made grammatical sense. If a semantic substitution that did not make grammatical sense was made, it was not counted as a gist point (see Example 1). Please refer to the section below for information regarding phrasal substitutions.
2. **Morphological substitutions** (e.g., inflectional changes) were given 1 point if they retained the overall meaning of the sentence and made grammatical sense. **Verb tense changes**, which fall within the category of morphological substitutions, were also given 1 point. See Example 2.
3. **Direct determiner substitutions** were given 1 point (see Example 4). **Pronoun substitutions** for **determiners** were also given 1 point (see Example 3). Pronoun substitutions for other pronouns were not given any points.
4. **Prepositional changes**, if they retained the meaning of the sentence, were given 1 point. See Example 2.

Phrasal substitutions that were given a varying number of points:

Importantly, phrasal substitutions were given the same number of points as the verbatim phrase would have received (e.g., if a phrase consisting of two words is substituted for one word, only 1 gist points was given). This was done to avoid inflating a score when a substitution contained more words than the target. See Example 4.

Examples (gist words are in **bold**):

Number	Type of Gist Word	Target Sentence (Number of Words)	Subject's Recall Attempt (Verbatim Score)	Additional Gist Points
1	Prepositional change, Semantic substitution	My under standing relied upon my being able to interpret the data easily. (12)	My understanding relied on my ability to interpret the data easily. (9)	2
2	Morphological substitution (verb tense)	The skydiver leapt from the plane, seeing dot-like trees on the ground below. (13)	Skydiver leapt from the plane and saw dot-like trees down ground below. (9)	1
3	Pronoun substitution for determiner	A peacock displays its beautiful plumage in order to attract a suitable mate. (13)	A peacock displays its beautiful plumage in order to attract its mate. (11)	1
4	Phrasal substitution	Computers can replace human workers in some but not all white collar jobs. (13)	Computers can replace workers in only a limited number of white collar jobs. (8)	4

Quantification of error types

For each recall attempt, we counted the occurrence of six types of errors.

1. Major order change
One major order change might include:
One open-class word exchanged for another, or;
One phrase or clause exchanged for another.
Example: "The ship deviated from its course during the night" → "The ship deviated in the night from its course."
2. Unrelated addition (open-class words)
An unrelated addition included any open-class word that was included in the recall attempt but did not occur in the target sentence, was *not* synonymous with a target word, and could not be considered as another category of error defined below.
Example: see #3 below.
3. Semantic substitution (open-class words)
A semantic substitution included any open-class word that was exchanged for another and *was* synonymous with the original word.

Example: “The stranger impressed the villagers with his honesty and integrity. → The stranger impressed the **town** with his **kindness** and **generosity**.”

Here, *town* is scored as a semantic substitution, as a near-synonym for *villagers*. *Kindness* and *generosity* are scored as unrelated additions, because they are not synonymous with the target words.

Grammatical substitution (closed-class words)

One grammatical substitution might include:

- One closed-class word exchanged for another, or;
- An original word that was pluralized, or;
- An original word that was changed in tense, or;
- Two words that were contracted.

Examples: “Literature attracts students who enjoy reading poems more than doing math problems.” → “Literature attracts students **that read** more than those that **play sports**.”

That and *read* are both scored as grammatical substitutions, while *play* and *sports* are unrelated substitutions.

Phonological substitution

A phonological substitution included any open-class word that was changed to another that sounded phonologically similar to the original word.

Example: “The train halted and a long line of passengers began disembarking onto the platform.” → “The train halted and there was a long line of passengers coming on the **plateau**.”

Plateau is a phonological substitution for *platform*. Other errors in the sentence are not discussed here.

Open-class omissions

An open-class omission included any complete omission of an open-class word.

A semantic substitution was not counted as an open-class omission, but an open-class phonological substitution was counted as an open-class omission.

Example: “The tire ruptured because of an enormous nail protruding from the road surface.” → “The tire ruptured due to the nail protruding from the road.”

The omissions of *enormous* and *surface* result in two errors being scored in this category.

Appendix D: Full sentence list

Concrete sentences

The waiter dropped many of the plates as he served the hungry customers.
Giraffes evolved to be tall enough to eat leaves on the upper branches of trees.
The thieves siphoned gasoline primarily from larger vehicles that held greater amounts.

The angry housewife locked the bedroom door so her husband could not enter.
The farmer woke early to collect the eggs before the summer sun became unbearable.

The audience gasped as the tightrope walker struggled to regain her balance.

The flowers wilted in the garden after the caretaker forgot to water them.
The lions descended upon the helpless wildebeest and subsisted on its meat for several days.

Rainbows have colored the sky this week during the ongoing thunderstorms.
The fog obscured the view of the city skyline from the expressway.

The cart rolled down the corridor and crashed into the overstuffed bookshelf.
The team enjoyed celebratory ice cream sandwiches on the field following the winning goal.

The teenager decorated her room but accidentally spilled paint on the bed.
The turtle retracted its head into its shell when a child tried to touch it.
The mirror shattered into several tiny pieces when I dropped it on the stone floor.

The archer launched the arrow directly into the center of the distant target.
The boy sneaked himself some chocolate-chip cookies while his mother was away.

The skydiver leapt from the plane, seeing dot-like trees on the ground below.
The boats evacuated the sinking battleship’s crew, saving hundreds of lives.
Cheetahs run faster than the antelopes and gazelles that they eat.

Swimmers dive off the starting blocks immediately when the whistle blows.
The snake handler wrapped his finger in a Band-Aid after getting a nasty bite.

The athlete jumped over the chainlink fence mainly to impress the cheerleaders.
The building emptied after some mischievous children set off the fire alarm.
The gorilla beat his chest as the school children watched from behind the glass.

Bears may attack hikers without warning when encountered in a wilderness setting.

The grandchildren removed their shoes just before stepping onto the expensive carpet.

The baby slept peacefully in the carseat throughout the entire car ride.

The tire ruptured because of an enormous nail protruding from the road surface.

Boxes are accumulating in the mailroom because the workers are on strike.

The referee declared the match over after the defending champion was knocked out.

The hostess ate all the leftover food after the party guests had all gone home.

The ship deviated from its course during the night and collided with an iceberg.

The football slipped out of the receiver’s hands, but was recovered by a defender.

A peacock displays its beautiful plumage in order to attract a suitable mate.
 Sausages can burn if left to sizzle on the barbecue for too long.
 A bridge crosses over the river, connecting the two halves of the city.
 The cookies rested on the granite countertop until they were sufficiently cool.
 The stereo shook the windows of the apartment and reverberated throughout the building.
 The golfer drove the ball onto the distant green with one excellent shot.
 The mechanic screamed when his helper accidentally dropped a wrench on his foot.
 A tomado destroyed the farmer's house and scattered the pieces over the fields.
 The bear hibernated in a cave all winter, emerging only when the snow had melted.
 The bleachers collapsed when the spectators all stomped their feet simultaneously.
 The juggler kept all three balls in the air while somersaulting off of her unicycle.
 Retrievers will sniff the ground obsessively until they discover the source of a smell.
 The delivery man waited at the door and ground his teeth in impatience.
 The potato chips crunched loudly in the teeth of the kids in the movie theatre.
 The nurses rushed into the room when the patient pushed the alarm button.
 The groom slid the ring onto her finger with an enormous smile on his face.

Abstract sentences

The commandments prohibit adultery as a severe offense against the community.
 The guidelines summarize past case outcomes and help to streamline future work.
 A resume will boost one's self-esteem by summarizing all of one's accomplishments.
 Distrust deepened into suspicion and ultimately ruined the deal between the men.
 The vision arose from a brainstorming session on a Friday afternoon.
 The group failed to achieve the stated objectives, much to their leader's disappointment.
 The structure of the company impairs attempts to improve efficiency.
 The actress feared that the shocking revelations would affect her career.
 Quotas are tightening as applications for admission surge to ever higher levels.
 The accountant regretted that he did not understand finance well enough to do the job.
 Abortion looms as a particularly divisive issue in the coming elections.
 The stranger impressed the villagers with his honesty and integrity.
 The manual includes instructions for keeping the software up to date.
 Income rises at about the same rate as inflation in a healthy economy.
 The urge subsided as quickly as it had come and she no longer wanted it.
 The formula calculates the likelihood of rain or snow over the next week.
 The party refused to support the new minister and demanded her resignation.
 The senator argued strongly in favor of the bill, until it was finally passed.
 Stocks have declined in value recently as confidence in the market has fallen.
 Software changes rapidly, making it hard to share projects between sites.
 A good parent teaches strong morals but can not force a child to adopt them.
 The bible clearly states the laws that believers must live by every day.

Yeast serves as an increasingly important model for the study of life.
 The health of older citizens has been improving under the current system.
 Literature attracts students who enjoy reading poems more than doing math problems.
 Game theory models situations in which outcomes are dependent on choices made by others.
 The scientist considered the problem but finally gave up on solving it.
 The patient complained of chronic unrelenting pain despite the medication.
 Tuition has escalated over the past few years despite an economic downfall.
 The loans came due at the end of the year, but they have not been repaid.
 The doctor retired because there was very little demand for his services.
 Transportation poses a major challenge to long-term planning in cities.
 Book themes provide the reader with an easy method to analyze classic literature.
 The internet distracts many students from their studies and harms productivity.
 Our friends discuss world politics much more frequently than most people do.
 The clerk forgot the procedure for reconciling the budget at the end of the year.
 Scandals plagued the party members and ended their hopes for re-election.
 This miracle inspired the townspeople to a higher level of faith.
 My understanding relied upon my being able to interpret the data easily.
 The journalist pondered his next career move following a disastrous job interview.
 The solution appears obvious if you spend enough time working on it.
 Some pundits ignore all information that does not fit with their position.
 The president handled the crushing defeat gracefully despite his bitter disappointment.
 The journal accepted the article after the revisions were submitted.
 Counselors advise students on their options when choosing a university.
 Leaders possess attributes that make others want to help them achieve their goal.
 Calculus ruins the career plans of students who would otherwise succeed.
 The voters expressed their preference for a vastly different style of government.
 The Greeks developed a complex mythology to explain the natural world.
 The afterlife varies greatly in importance across different religions of the world.
 The technique assists many students in the tedious chore of memorizing the tables.
 Placebos can cure many diseases if their secrecy is maintained effectively.
 The warning affected everyone's mood as we tried to continue with our normal work.
 Good employers create opportunities that allow their employees to advance steadily.
 The official said that he was not authorized to change the decision.
 The government suggested that everyone over the age of eighteen should vote.
 The policy solved some of our biggest problems but also created new ones.
 The pianist evaluated his prospects before switching majors to business.
 The lawyer spent hours preparing for the trial before the case was settled.
 The results implied that the theory had been misguided from the very beginning.
 Some citizens violate civil laws without knowing that they are doing so.
 Celebrities support a large media industry simply by living their lives.
 Jealousy can threaten friendships when one person is more successful than the other.
 Ideas translate best into action when the motivation is there to see them through.
 An advisory was issued for all man-made products containing sulphuric acid.
 Education can lift people up into a higher social status than their parents achieved.

The store conducted most of its business during the holiday season.
 Our city council assembled only when there were important issues to be discussed.
 Time passed by slowly that day as we eagerly anticipated the evening's events.
 The setback crushed the workers' hopes for a larger annual bonus.
 Computers can replace human workers in some but not all white collar jobs.
 Energy moves through the food chain mainly in chemical form.
 Disagreements over details delayed the approval of the constitution for years.
 Most theories build upon existing ones by expanding as we discover new information.
 The message contained explicit instructions from the chief on how to proceed.
 The article reminded the girl of several enjoyable conversations with her father.
 The future of the country lies in the able hands of the younger generation.
 The statement read that the citizens would no longer need such tight rations.
 Some people aspire to achieve greatness, while others are happy with the status quo.
 Newlyweds always look forward to a happy marriage free of major tensions.

References

- Aldridge, J. W., & Crisp, T. (1982). Maintenance rehearsal and long-term recall with a minimal number of items. *American Journal of Psychology*, 95, 565–570. doi:10.2307/1422187
- Alloway, T. P. (2007). Investigating the roles of phonological and semantic memory in sentence recall. *Memory*, 15, 605–615. doi:10.1080/09658210701450877
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417–423. doi:10.1016/S1364-6613(00)01538-2
- Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, 36, 189–208. doi:10.1016/S0021-9924(03)00019-4
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press.
- Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14, 575–589. doi:10.1016/S0022-5371(75)80045-4
- Baddeley, A. D., Vallar, G., & Wilson, B. (1987). Sentence comprehension and phonological memory: Some neuropsychological evidence. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 509–529). Hillsdale, NJ: Erlbaum.
- Baldo, J. V., Klosternann, E. C., & Dronkers, N. F. (2008). It's either a cook or a baker: Patients with conduction aphasia get the gist but lose the trace. *Brain and Language*, 105, 134–140. doi:10.1016/j.bandl.2007.12.007
- Belleville, S., Caza, N., & Peretz, I. (2003). A neuropsychological argument for a processing view of memory. *Journal of Memory and Language*, 48, 686–703.
- Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19, 2767–2796. doi:10.1093/cercor/bhp055
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Cambridge, MA: O'Reilly.
- Bireta, T. J., Neath, I., & Surprenant, A. M. (2006). The syllable-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, 13, 434–438. doi:10.3758/BF03193866
- Bourassa, D. C., & Besner, D. (1994). Beyond the articulatory loop: A semantic contribution to serial order recall of subspan lists. *Psychonomic Bulletin & Review*, 1, 122–125.
- Braver, T. S., Barch, D. M., Kelley, W. M., Buckner, R. L., Cohen, N. J., Miezin, F. M., & Petersen, S. E. (2001). Direct comparison of prefrontal cortex regions engaged by working and long-term memory tasks. *NeuroImage*, 14, 48–59.
- Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology*, 26, 467–482.
- Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1998). Making memories: Brain activity that predicts how well visual experience will be remembered. *Science*, 281, 1185–1187. doi:10.1126/science.281.5380.1185
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904–911. doi:10.3758/s13428-013-0403-5
- Butterworth, B., Shallice, T., & Watson, F. L. (1990). Short-term retention without short-term memory. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 187–214). Cambridge, UK: Cambridge University Press.
- Cabeza, R., Dolcos, F., Graham, R., & Nyberg, L. (2002). Similarities and differences in the neural correlates of episodic memory retrieval and working memory. *NeuroImage*, 16, 317–330.
- Cameron, K. A., Haarmann, H. J., Grafman, J., & Ruchkin, D. S. (2005). Long-term memory is the representational basis for semantic verbal short-term memory. *Psychophysiology*, 42, 643–653. doi:10.1111/j.1469-8986.2005.00357.x
- Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13, 53–76.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290. doi:10.1037/1040-3590.6.4.284
- Clark, H. H. (1973). The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359. doi:10.1016/S0022-5371(73)80014-3
- Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.
- Clarkson, P., & Rosenfeld, R. (1997, September). *Statistical language modeling using the CMU-Cambridge toolkit*. Paper presented at Eurospeech 97, Rhodes, Greece.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684. doi:10.1016/S0022-5371(72)80001-X
- D'Agostino, P. R., O'Neill, B. J., & Paivio, A. (1977). Memory for pictures and words as a function of level of processing: Depth or dual coding? *Memory & Cognition*, 5, 252–256.
- Dark, V. J., & Loftus, G. R. (1976). The role of rehearsal in long-term memory performance. *Journal of Verbal Learning and Verbal Behavior*, 15, 479–490.
- Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, 112, 3–42. doi:10.1037/0033-295X.112.1.3
- Davelaar, E. J., Haarmann, H. J., Goshen-Gottstein, Y., & Usher, M. (2006). Semantic similarity dissociates short- from long-term recency effects: Testing a neurocomputational model of list memory. *Memory & Cognition*, 34, 323–334. doi:10.3758/BF03193410

- Erdelyi, M. H. (2010). The ups and downs of memory. *American Psychologist*, 65, 623–633. doi:10.1037/a0020440
- Haarmann, H. J., Davelaar, E. J., & Usher, M. (2003). Individual differences in semantic short-term memory capacity and reading comprehension. *Journal of Memory and Language*, 48, 320–345.
- Haarmann, H., & Usher, M. (2001). Maintenance of semantic information in capacity-limited item short-term memory. *Psychonomic Bulletin & Review*, 8, 568–578.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8, 23–34.
- Hamilton, A. C., Martin, R. C., & Burton, P. C. (2009). Converging functional magnetic resonance imaging evidence for a role of the left inferior frontal lobe in semantic retention during language comprehension. *Cognitive Neuropsychology*, 26, 685–704. doi:10.1080/02643291003665688
- Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, 30, 685–701.
- Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1217–1232. doi:10.1037/0278-7393.23.5.1217
- Ide, N., & MacLeod, C. (2001, March). *The American National Corpus: A standardized resource of American English*. Paper presented at the Corpus Linguistics 2001 Conference, Lancaster, UK.
- Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 336–358. doi:10.1037/0278-7393.26.2.336
- Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage*, 54, 2446–2461. doi:10.1016/j.neuroimage.2010.09.045
- Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In E. W. Hinrichs & D. Roth (Eds.), *ACL '03: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 423–430). New York, NY: ACM Press. doi:10.3115/1075096.1075150
- Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19, 313–330.
- Martin, R. C., & He, T. (2004). Semantic short-term memory and its role in sentence processing: A replication. *Brain and Language*, 89, 76–82. doi:10.1016/s0093-934x(03)00300-6
- Martin, N., & Saffran, E. (1997). Language and auditory-verbal short-term memory impairments: Evidence for common underlying processes. *Cognitive Neuropsychology*, 14, 641–682.
- Martin, R. C., Shelton, J. R., & Yaffee, L. S. (1994). Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. *Journal of Memory and Language*, 33, 83–111.
- McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language*, 58, 480–494. doi:10.1016/j.jml.2007.04.004
- McCarthy, R., & Warrington, E. K. (1984). A two-route model of speech production: Evidence from aphasia. *Brain*, 107, 463–485.
- McCarthy, R. A., & Warrington, E. K. (1987). The double dissociation of short-term memory for lists and sentences. Evidence from aphasia. *Brain*, 110, 1545–1563.
- Miller, G. A., & Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. *American Journal of Psychology*, 63, 176–185.
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4, 61–64. doi:10.3758/s13414-012-0291-2
- Paivio, A., Clark, J. M., & Khan, M. (1988). Effects of concreteness and semantic relatedness on composite imagery ratings and cued recall. *Memory & Cognition*, 16, 422–430. doi:10.3758/BF03214222
- Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8, 976–987.
- Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, 29, 633–654.
- Race, E., Palombo, D. J., Cadden, M., Burke, K., & Verfaellie, M. (2015). Memory integration in amnesia: Prior knowledge supports verbal short-term memory. *Neuropsychologia*, 70, 272–280. doi:10.1016/j.neuropsychologia.2015.02.004
- Ranganath, C., & Blumenfeld, R. S. (2005). Doubts about double dissociations between short- and long-term memory. *Trends in Cognitive Sciences*, 9, 374–380. doi:10.1016/j.tics.2005.06.009
- Reilly, J., & Kean, J. (2007). Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cognitive Science*, 31, 157–168.
- Richardson, J. T. E. (1975). Concreteness and imageability. *Quarterly Journal of Experimental Psychology*, 27, 235–249. doi:10.1080/14640747508400483
- Romani, C., McAlpine, S., & Martin, R. C. (2008). Concreteness effects in different tasks: Implications for models of short-term memory. *Quarterly Journal of Experimental Psychology*, 61, 292–323. doi:10.1080/17470210601147747
- Rose, N. S. (2013). Individual differences in working memory, secondary memory, and fluid intelligence: Evidence from the levels-of-processing span task. *Canadian Journal of Experimental Psychology*, 67, 260–270. doi:10.1037/a0034351
- Rose, N. S., Buchsbaum, B. R., & Craik, F. I. M. (2014). Short-term retention of a single word relies on retrieval from long-term memory when both rehearsal and refreshing are disrupted. *Memory & Cognition*, 42, 689–700. doi:10.3758/s13421-014-0398-x
- Rose, N. S., & Craik, F. I. M. (2012). A processing approach to the working memory/long-term memory distinction: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1019–1029. doi:10.1037/a0026976
- Rose, N. S., Craik, F. I., & Buchsbaum, B. R. (2015). Levels of processing in working memory: Differential involvement of frontotemporal networks. *Journal of Cognitive Neuroscience*, 27, 522–532. doi:10.1162/jocn_a_00738
- Rose, N. S., Myerson, J., Roediger, H. L., III, & Hale, S. (2010). Similarities and differences between working memory and long-term memory: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 471–483. doi:10.1037/a0018405
- Rose, N. S., Olsen, R. K., Craik, F. I., & Rosenbaum, R. S. (2012). Working memory and amnesia: The role of stimulus novelty. *Neuropsychologia*, 50, 11–18. doi:10.1016/j.neuropsychologia.2011.10.016
- Ruchkin, D. S., Grafman, J., Cameron, K., & Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral and Brain Sciences*, 26, 709–728. **disc. 728–777.**
- Rummer, R., & Engelkamp, J. (2001). Phonological information contributes to short-term recall of auditorily presented sentences. *Journal of Memory and Language*, 45, 451–467. doi:10.1006/jmla.2000.2788
- Rummer, R., & Engelkamp, J. (2003). Phonological information in immediate and delayed sentence recall. *Quarterly Journal of Experimental Psychology*, 56A, 83–95. doi:10.1080/02724980244000279

- Rundus, D. (1977). Maintenance rehearsal and single-level processing. *Journal of Verbal Learning and Verbal Behavior*, 16, 665–681.
- Sampson, G. (1997). Depth in English grammar. *Journal of Linguistics*, 33, 131–151.
- Schweppe, J., Rummer, R., Bormann, T., & Martin, R. C. (2011). Semantic and phonological information in sentence recall: Converging psycholinguistic and neuropsychological evidence. *Cognitive Neuropsychology*, 28, 521–545. doi:[10.1080/02643294.2012.689759](https://doi.org/10.1080/02643294.2012.689759)
- Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology*, 51A, 283–304. doi:[10.1080/713755759](https://doi.org/10.1080/713755759)
- Shivde, G., & Anderson, M. C. (2011). On the existence of semantic working memory: Evidence for direct semantic maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1342–1370. doi:[10.1037/a0024832](https://doi.org/10.1037/a0024832)
- Shivde, G., & Thompson-Schill, S. L. (2004). Dissociating semantic and phonological maintenance using fMRI. *Cognitive, Affective, & Behavioral Neuroscience*, 4, 10–19. doi:[10.3758/CABN.4.1.10](https://doi.org/10.3758/CABN.4.1.10)
- Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, 52, 452–473.
- Surprenant, A. M., & Neath, I. (2008). The 9 lives of short-term memory. In A. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 16–43). Hove, UK: Psychology Press.
- Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Quarterly*, 30, 415–433.
- Wagner, A. D., Koutstaal, W., & Schacter, D. L. (1999). When encoding yields remembering: Insights from event-related neuroimaging. *Philosophical Transactions of the Royal Society B*, 354, 1307–1324.
- Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104, 444–466.