Psychon Bull Rev
DOI 10.3758/s13423-015-0840-5

BRIEF REPORT

Dynamic memory searches: Selective output interference for the memory of facts

William R. Aue[1,3] · Amy H. Criss[1] · Melissa A. Prince[2]

Abstract The benefits of testing on later memory performance are well documented; however, the manner in which testing harms memory performance is less well understood. This research is concerned with the finding that accuracy decreases over the course of testing, a phenomena termed "output interference" (OI). OI has primarily been investigated with episodic memory, but there is limited research investigating OI in measures of semantic memory (i.e., knowledge). In the current study, participants were twice tested for their knowledge of factual questions; they received corrective feedback during the first test. No OI was observed during the first test, when participants presumably searched semantic memory to answer the general-knowledge questions. During the second test, OI was observed. Conditional analyses of Test 2 performance revealed that OI was largely isolated to questions answered incorrectly during Test 1. These were questions for which participants needed to rely on recent experience (i.e., the feedback in episodic memory) to respond correctly. One possible explanation is that episodic memory is more susceptible to the sort of interference generated during testing (e.g., gradual changes in context, encoding/updating of items) relative to semantic memory. Alternative explanations are considered.

Keywords Interference · Knowledge · Recognition · Testing

✉ William R. Aue
william.aue@gmail.com

[1] Department of Psychology, Syracuse University, Syracuse, NY 13244, USA

[2] School of Psychology, University of Sydney, Sydney, NSW 2006, Australia

[3] Present address: Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

Understanding the value of testing and the use of testing as a learning device is critically important, especially with the growing use of high-stakes testing in education. What makes testing benefits particularly interesting is the differential impact of information learned during testing relative to information learned during an equivalent amount of time spent studying the material. For example, participants tend to remember information better if they complete a free recall test than if they simply study the material again (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a). While testing has robust and reliable benefits (see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Roediger & Karpicke, 2006b, for reviews), there are also negative consequences of testing (see Malmberg, Lehman, Annis, Criss, & Shiffrin, 2014). For instance, retrieving a subset of studied items associated with a particular cue (e.g., category membership) during test can impair retrieval of other items from the same set (Malmberg, Criss, Gangwani, & Shiffrin, 2012; Roediger, 1973; Slamecka, 1968). A substantial negative consequence of testing is the finding that performance decreases over the course of a test list (Criss, Malmberg, & Shiffrin, 2011; Murdock & Anderson, 1975; Ratcliff & Hockley, 1980; Roediger, 1974; Roediger & Schmidt, 1980). This finding, termed "output interference" (OI), has been modeled as the result of encoding during test by either updating existing memories with information gained during the test or adding new traces to episodic memory. Consequently, items that are tested toward the end of a test list suffer from additional interference generated by the information added to memory during the course of testing (Criss et al., 2011).

The negative effects of OI are robust and widespread. The effect has been observed in both recall (Raaijmakers & Shiffrin, 1981; Roediger & Schmidt, 1980) and recognition (Criss et al., 2011; Ratcliff & Hockley, 1980; Annis, Malmberg, Criss, & Shiffrin, 2013; Malmberg et al., 2012) and with different methodological and stimulus variations, including categorized (Criss et al., submitted; Malmberg

Published online: 25 April 2015

et al., 2012; Roediger & Schmidt, 1980) and randomly generated lists of words (Criss et al., 2011; Annis et al., 2013; Malmberg et al., 2012), across study test lag after both short and long delays (Criss et al., 2011; Roediger & Schmidt, 1980; A. D. Smith, 1973) for target- and distractor-free tests (Koop, Criss, & Malmberg, 2015) and following feedback (Criss et al., 2011; Koop et al, 2015). Additionally, the more information added to memory during testing (e.g., a four alternatives test vs. two), the greater the magnitude of OI (Murdock & Anderson, 1975). However, OI can be attenuated under certain test circumstances. For instance, switching between categories of items during testing (e.g., testing all items from one category followed by all items from another category) demonstrates a buildup of OI within a category but a release from OI following a category switch mid-test (Criss et al., submitted; Malmberg et al., 2012), similar to a release from proactive interference (e.g., Wickens, Born, & Allen, 1963).

Interestingly, it is not simply engaging in a task or the presence of items that causes OI; adding a semantic memory task during recognition testing does not contribute to OI. Annis et al. (2013) for example, tested participants for their memory of a studied list of words. Forced-choice recognition test trials were interleaved with either a lexical decision (LD) task, where participants simply had to decide whether a presented set of letters was a word or a nonword; a gender identification task, where participants had to determine the gender of a presented face; or no task at all. While OI was observed during recognition testing, it was not differentially impacted by the presence (e.g., task versus no-task) or nature of the intervening task (e.g., LD vs. gender ID). This suggests that it is the act of episodic retrieval that is driving OI.

The Annis et al. (2013) data demonstrate that retrieving from semantic traces does not add to the OI measured for an episodic task alone. However, it says nothing about OI occurring within a semantic task. Although the semantic and episodic memory systems dissociate in a variety of functional, pharmacological, and structural ways, the systems are heavily dependent on one another (Greenberg & Verfaellie, 2010; Nyberg & Tulving, 1996; Tulving, 1972). Semantic information contributes to the retrieval of episodic information (e.g., Howard & Kahana, 2002; Prince, Tsukiura, & Cabeza, 2007) and episodic experience primes performance on semantic tasks (Schooler, Shiffrin, & Raaijmakers, 2001; Jacoby & Dallas, 1981; Ratcliff & McKoon, 1997; Reder, Park, & Kieffaber, 2009). Moreover, semantic knowledge is necessarily episodic at one point (i.e., when the fact was initially learned; A. B. Nelson & Shiffrin, 2013; Mueller & Shiffrin, 2006; Schooler et al., 2001); however, through repeated exposure the information becomes more complete and decontextualized. The result is a more durable, semantic memory resistant to episodic interference (Tulving, 1972).

Our aim was to better understand the nature of OI in relation to semantic and episodic memory. Specifically, we examined whether performance in a semantic task (e.g., test of knowledge) would suffer from OI by examining performance on a test of general knowledge questions. Participants answered the same general knowledge questions before (Test 1) and after (Test 2) receiving corrective feedback. In Test 1, participants were presumably searching semantic knowledge to answer the questions given that there is likely to be no recent or specific episodic memory for the tested items. However, Test 2 has the potential to be a more episodic task given the corrective feedback received during Test 1.

Method

Participants Sixty-six members of the Syracuse University research pool participated in the experiment for course credit.

Materials The stimuli used for the experiment were 300 general knowledge questions developed by T. O. Nelson and Narens (1980) and revised by Tauber, Dunlosky, Rawson, Rhodes, and Sitzman (2013). The materials were adapted for use in a four-alternative forced choice recognition (4AFC) procedure by generating plausible foil items for each target response. Foil items were chosen based on semantic proximity to the target response as measured by latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998).[1] Reasonable responses that were similar parts of speech, were not synonyms, were not another form of the response, and did not appear elsewhere in the stimulus set were selected. In instances where one or more foil responses were unable to be retrieved from LSA we used a simple internet search of the question and selected a response from amongst the top results according to the same constraints. The experiment was administered using MATLAB (R2011a) and Psychtoolbox v3.0 (Brainard, 1997).

Design and procedure During the experiment, participants were tested on a total of 153 questions. There were 150 critical questions during Test 1, followed by three buffer questions that appeared only on Test 1. Although both lists were randomized, the buffer questions ensured that the final questions of Test 1 were not the first questions of Test 2. The position of target and foil response options was also randomized for both tests. Testing was self-paced. Tests 1 and 2 differed in only a few details. The first difference is that during Test 1 participants received corrective feedback for their response whereas no feedback was provided during Test 2. In Test 1, feedback was provided immediately following the response, centered on the screen, and remained on the screen for 1.5 s. If the response was correct, participants were told as much. If the response was incorrect, they were told so and provided with

[1] http://lsa.colorado.edu/

the correct answer. The second difference was that participants received different instructions during Test 1 and Test 2. As in Tauber et al. (2013), participants were encouraged to take their time and answer the questions to the best of their ability during Test 1. However, participants were not informed that they were going to be tested on the material again. On Test 2, participants were told they would be tested again on the Test 1 material and were encouraged to consider their Test 1 response and feedback when answering the question. Between tests participants completed a 60 s arithmetic task where they kept a running summation of a series of single digits. The entire experiment lasted approximately 60 min. A depiction of the experimental design and three sample questions are provided in Fig. 1.

Data analysis Test performance was analyzed using a Bayesian multiple linear regression developed by Kruschke (2011) and cross-validated with a frequentist regression. Interpretation of the Bayesian model parameters is identical to that of a typical regression from the frequentist tradition. We chose this approach because it allows us to quantify the amount of evidence in favor or against a null effect (e.g., Rouder, Speckman, Dongchu, Morey, & Iverson, 2009). The focus of our analysis was the change in performance across test trials, specifically whether there is a nonzero slope for the performance change. Moreover, we examined whether the slope of performance across trials changed from Test 1 to Test 2.

For each participant, the 150 test trials were binned into 10 blocks of 15 trials each. Entered into the regression model were standardized block-level performance data, block number (1:10), dummy-coded test number (Test 1 = 0, Test 2 = 1), and the interaction of block number and test number.

We used an identical approach for the conditional analysis of the Test 2 performance. In this case, Test 2 performance was binned into six blocks of 25 trials each, separately for trials that were either correct or incorrect during Test 1. For the regression model, standardized block-level performance data, block number (1:6), dummy-coded Test 1 accuracy (Test 1 correct = 1, Test 1 incorrect = 0), and the interaction of block number and Test 1 accuracy condition.

For the Bayesian analyses, we adopted the same uninformed priors and procedural details recommended by Kruschke (2011).[2] The credibility of the results was evaluated by examining whether the 95 % highest density interval (HDI) for the posterior slope parameter estimate included zero as a credible value. We also evaluated the level of evidence for the hypothesis that the slope was nonzero by estimating the Bayes Factor for the slope estimates. The Bayes Factor is a measure of the weight of evidence derived from the observed data (Jeffreys, 1961) and was estimated using a Savage-Dickey analysis (see Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Lee & Wagenmakers, 2014, for helpful tutorials).

For the Savage-Dickey analysis, we simplified the analysis into two competing hypotheses regarding the slope, or the difference in slope in the case of the interactions, of performance change across test block:

1. The null hypothesis that OI was not present and that the slope is 0 (H_0: β = 0).
2. The alternative hypothesis that the slope is nonzero (H_1: β ≠ 0).

The Savage-Dickey estimation of the Bayes Factor (B_{01}) is the ratio of the likelihood of observing a zero slope in the posterior distribution relative to observing a zero slope in the prior distribution. A BF of less than 1 indicates that observing a slope of zero is more likely in the prior distribution and provides greater evidence in favor of the null hypothesis that OI is not present and that slope of the observed data is zero (i.e., β = 0). A BF greater than 1 indicates evidence in favor of the alternative hypothesis (i.e., $\beta_1 \neq 0$), that the slope of the observed data is credibly nonzero (i.e., β ≠ 0). The analyses were completed in R (v2.15; R Core Team, 2013) using the JAGS software (v3.1.0) and "rjags" package (Plummer, 2013) for R.

Results

The descriptive data and Bayesian modeling results for Test 1 and Test 2 are presented in Fig. 2. Visually, the data suggest that performance on Test 1 does not change across successive trials, whereas performance on Test 2 decreases across trial. The profile plot in Fig. 2a depicts performance for a given test block (containing 15 test trials each) of Test 1. Indeed, the interaction term, representing the difference in slope of the Test 1 and 2 data, was negative indicating a bigger value for Test 2 (β = -.007, 95 % HDI: -.013, -.0004) and did not include zero as a credible value (BF_{01} = .341). Therefore, there is very strong evidence that the Test 2 performance declines over the test block and Test 1 does not change, with modest support for a difference between the slopes of Test 1 and Test 2.

Examining the simple slopes of Test 1 and Test 2 separately, the slope for Test 1 was effectively flat (β = -0.0006, 95 % HDI: -.005, .003) and has an HDI that encompasses zero as a

[2] Model runs included four chains with 50,000 iterations, each with 1,000 burn-in steps and thinned at every 50th iteration. Convergence of the chains for the parameters of interest was confirmed with a Gelman-Rubin ratio ($\hat{R}$) of 1. For the model the prior for each predictor (β) was normally distributed with a mean of 0 and standard deviation of .01.

Fig. 1 A depiction of the experimental design employed and three sample questions that participants might have experienced during the experiment

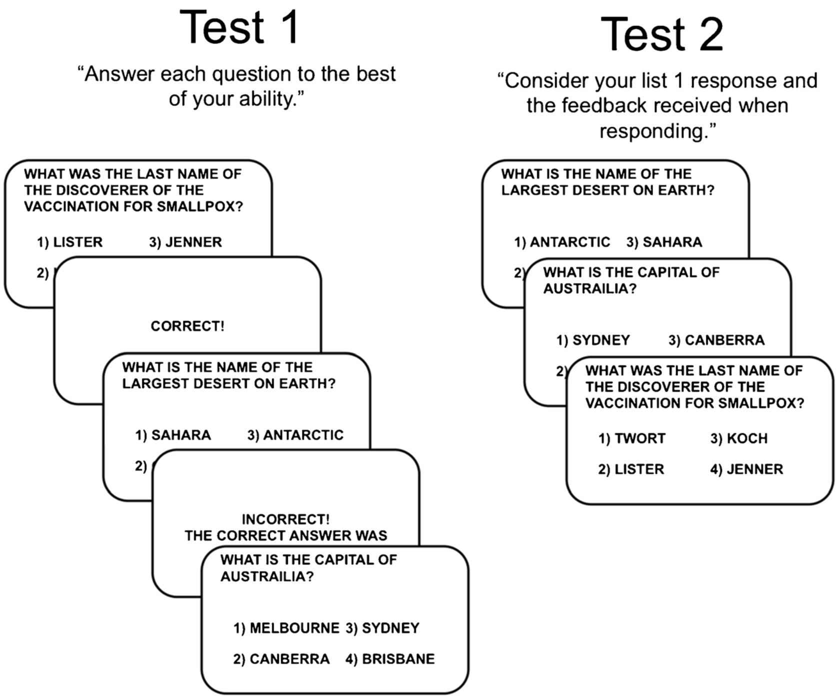
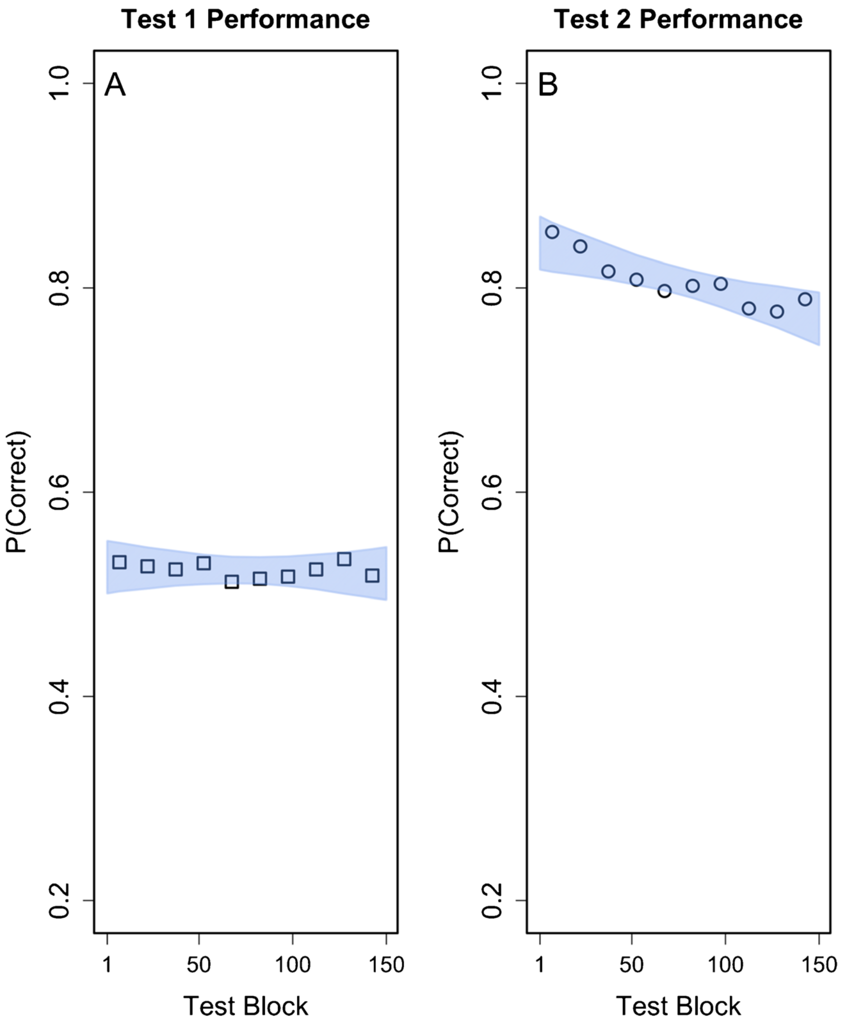

Test 1
"Answer each question to the best of your ability."
WHAT WAS THE LAST NAME OF THE DISCOVERER OF THE VACCINATION FOR SMALLPOX?
1) LISTER
3) JENNER
2)
CORRECT!
WHAT IS THE NAME OF THE LARGEST DESERT ON EARTH?
1) SAHARA
3) ANTARCTIC
2)
INCORRECT!
THE CORRECT ANSWER WAS
WHAT IS THE CAPITAL OF AUSTRAILIA?
1) MELBOURNE
3) SYDNEY
2) CANBERRA
4) BRISBANE
Test 2
"Consider your list 1 response and the feedback received when responding."
WHAT IS THE NAME OF THE LARGEST DESERT ON EARTH?
1) ANTARCTIC
3) SAHARA
2)
WHAT IS THE CAPITAL OF AUSTRAILIA?
1) SYDNEY
3) CANBERRA
2)
WHAT WAS THE LAST NAME OF THE DISCOVERER OF THE VACCINATION FOR SMALLPOX?
1) TWORT
3) KOCH
2) LISTER
4) JENNER

Fig. 2 Performance for Test 1 (**a**) and Test 2 (**b**). For each participant performance was averaged into 10 blocks of 15 trials each. The shaded region in Fig. 1 represents the 95 % highest density interval (HDI) of the posterior distribution of the predicted values and should be interpreted similarly to confidence intervals. Notably, no output interference is observed during Test 1, evidenced by the fact that performance does not change during the test. However, OI is observed during Test 2, where performance decreases monotonically during the test



credible value. Moreover, a zero value for slope is 4.2 times more likely for Test 1 than the alternative (BF_{01} = 4.21). For Test 2, the slope did not contain zero as a credible value (β = -.007, 95 % HDI: -.012, -.003). Furthermore, a nonzero slope is 40 times more likely than the null hypothesis (BF_{01} = .025).

A frequentist multiple regression performed on the data aggregated across subjects by block and test number corroborate the above results. Test Block and Test Number explained a significant amount of variance in performance, $F(3, 16) = 1, 441$, $p < .001$, $R^2 = .996$, $R^2_{\text{Adjusted}} = .996$). Moreover, the interaction of Test Block and Test Number was significant, β = -.006, $t(16) = -4.50$, $p < .001$. The slope of the change in performance across Test Block for Test 1 (β = -.0006) was shallower relative to the change in slope observed for Test 2 (β = -.007).

One possible explanation for the OI observed in Test 2 is that performance is a combination of searches of knowledge and recent memory, with the latter being more influenced by OI. To explore this possibility, we evaluated Test 2 performance conditionalized on the accuracy of the Test 1 response. If the Test 1 response was correct, then it is possible that the participant found the answer by a searching knowledge or, in rare cases, by correctly guessing. Under these conditions, OI may not be observed. If the Test 1 response was incorrect, participants may generate a correct Test 2 response on occasion by correctly guessing or by engaging a successful search of semantic memory that failed during Test 1. However, it seems reasonable that participants may also be relying on episodic memory for the Test 1 feedback to generate a correct Test 2 response, particularly given much higher accuracy for Test 2 than Test 1. Under these circumstances, OI should be observed.

As can be seen in Fig. 3a, when participants answered a question correctly during Test 1, performance on Test 2 did not decline across test position. However, when the question was incorrect on Test 1, performance on Test 2 decreased substantially across test position (see Fig. 3b). These patterns are substantiated by the Bayesian analysis. Indeed, the interaction term, representing the difference in slope of Test 2 when the Test 1 response was correct or incorrect, was credibly positive indicating a steeper slope for the latter (β = .016, 95 % HDI: .0006, .031) and did not include zero as a credible value (BF_{01} = .148). Examining the simple slopes, the slope for Test 2 given the Test 1 response was correct was relatively flat (β = -.006, 95 % HDI: -.016, .005), and has an HDI that encompasses zero as a credible value. However, the evidence in favor of a zero value is ambiguous because it is nearly equal to evidence in favor of a nonzero slope (BF_{01} = 1.04). Thus, we can neither reject nor accept the null hypothesis in this instance.

When the Test 1 response was incorrect, the decreasing slope during Test 2 was more pronounced (β = -.022, 95 % HDI: -.033, -.011) and did not contain zero as a credible value. In this case a nonzero value for slope is 1,000 times more likely than the null hypothesis (B_{01} = .001), indicating extreme evidence in favor of the alternative hypothesis.

The Bayesian results were again substantiated by a frequentist multiple regression on the data aggregated across participant by Test Block and Test 1 accuracy indicating that Test Block and Test 1 accuracy explained a significant amount of variance in performance, $F(3, 8) = 96.41$, $p < .001$, $R^2 = .973$, $R^2_{\text{Adjusted}} = .963$). Moreover, the interaction of Test Bin and Previous Response was significant, β = .016, $t(8) = 3.83$, $p = .005$. The slope of the change in performance across Test Block for Test 2 was shallower when the Test 1 response was correct (β = -.006) relative to when the Test 1 response was incorrect (β = -.022).

Discussion

In the current data, we observed that the pattern of output interference (OI) changed depending on whether it was the first or second time the set of questions had been answered. Specifically, no OI was observed during the initial test, but a robust OI effect was observed during the second test – after participants had received corrective feedback. Moreover, this OI was largely restricted to those questions for which the Test 1 response was incorrect. During the initial test, in the absence of recent episodic experience with the factual information, it is reasonable to presume that participants are searching knowledge (i.e., semantic memory), yet we observed no OI for Test 1. Participants could have recent experience with some of the information and, as a result, search episodic memory during Test 1, but such an occurrence would likely not be systematic. Critically, no OI was observed for Test 1. During the second test, it is reasonable to presume that participants are likely answering questions using a combination of knowledge and recent experience. For questions that were answered correctly during Test 1, participants could rely on knowledge as they presumably did during Test 1, to retrieve the correct answer. Of course, they could also search episodic memory. For questions that were answered incorrectly on Test 1, participants could rely on the episodic memory of the corrective feedback from Test 1 in order to answer Test 2. Of course they could also search knowledge, thought this is unlikely to lead to a correct answer given that it failed during Test 1. Importantly, Test 2 demonstrated a robust OI effect that was largely isolated to questions answered incorrectly during Test 1. We suggest that this indicates that OI is most pronounced for episodic memory.

Many models have attempted to connect the development of episodic memories and knowledge within a single framework (e.g., McClelland, McNaughton, & O'Reilly, 1995; Mueller & Shiffrin, 2006; A. B. Nelson & Shiffrin, 2013; Schooler et al., 2001; Shiffrin & Steyvers, 1997). The data

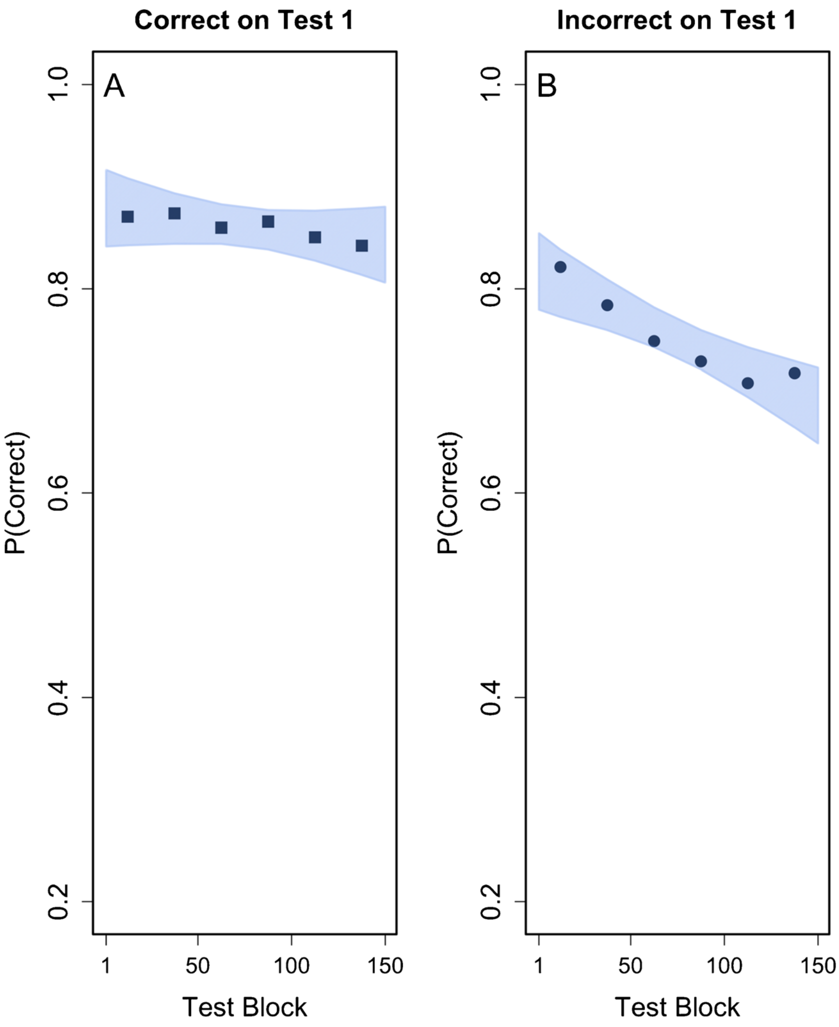


Fig. 3 Test 2 performance conditionalized on whether a given question was answered correctly (Panel **a**) or incorrectly (Panel **b**) on Test 1. Performance is averaged over 25 trials to generate test blocks. The shaded region in represents the 95 % highest density interval (HDI) of the values. No output interference was observed for Test 2 questions where the question was answered correctly during Test 1 (**a**). However, when a participant answered a question incorrectly on Test 1, a robust output interference effect was observed for Test 2 (**b**)

presented here are generally consistent with such frameworks as follows. Across the course of development and the learning of a new set of knowledge, information is added to semantic memory. Once that information is well learned, the semantic traces remain stable and can be accessed without affecting the representation. In contrast, information is continuously stored in episodic traces, and such traces are incomplete. Episodic traces, by definition, contain contextual information about the episode, whereas semantic traces are decontextualized. Retrieval from memory depends on the cue; specifically, the activation of episodic traces depends on the inclusion of contextual details in the cue. The cue match serves to narrow the set of retrieved traces to the relatively small set of those from the specific episode under investigation. Accordingly, the addition of information to these episodic traces or the addition of new episodic traces creates interference and harms performance in an episodic task. No such interference takes place with semantic retrieval because semantic traces are context-independent.

Several studies show what appears to be OI in free recall from semantic memory, in the form of slower retrieval after multiple retrieval attempts (e.g., Bousfield & Sedgewick, 1944). For example, in Blaxton and Neely (1983), participants were slower to generate a category exemplar when they had previously generated several exemplars from the same category relative to when they had generated only one member from the category. Response time as a function of trial is similar to that observed in retrieval from episodic memory in free recall (e.g., Rohrer & Wixted, 1994), potentially leading to the conclusion that OI in episodic and semantic memory is comparable. Two possible explanations can resolve the apparent discrepancy between these data and our data showing no OI in retrieval from semantic memory. First, as suggested by Bousfield, Sedgewick, and Cohen (1954) and subsequently became the basis for one explanation of the part list cuing effect (Raaijmakers & Shiffrin, 1980), retrieving an item from memory makes it more likely to be sampled again. In terms of the framework we describe above, retrieving from semantic memory causes storage of an episodic memory trace containing the retrieved item and current context. In other words, the observed interference in semantic retrieval is due to retrieval from episodic memory, much like the explanation for the current data. The benefit of the current experimental design is that it allows us to separately evaluate those trials that seemingly rely on episodic retrieval (incorrect on Test 1) from those trials that seemingly rely on retrieval from knowledge (correct on Test 1). A second possibility is that retrieval from semantic memory follows an optimal foraging strategy, switching between local and global cues. The slowing of response times could reflect the transitions between these types of cues (Hills, Jones, & Todd, 2012). This possibility attributes the slowing to a mechanism that is not related to OI and does not speak to our hypothesis of OI due to context-bound episodic retrieval or the absence of OI in context-independent semantic retrieval.

Alternative explanations based on methodology are also possible. For example, difficult retrieval could make a task susceptible to OI, and, indeed, the questions used were challenging. To investigate this we calculated the difficulty of each question by aggregating performance for the question across test position. We found that our randomization was successful in that question difficulty was evenly distributed across test position for Test 2. In contrast, there may be concern about the ability to detect OI due to a floor effect given that performance was worse during Test 1 relative to Test 2. However, this seems unlikely given that performance is well above chance and that OI has been reported even when performance is very low (e.g., Criss et al., 2011; Koop et al., 2015; Murdock & Anderson, 1975). Finally, suppose the critical factor is the similarity among to-be-retrieved items.[3] When the search set is similar, OI is likely to be observed and when it is dissimilar, OI is not likely to be observed. In the semantic retrieval studies described above, items shared categorical information whereas in our data the items shared the context of Test 1. On the surface, this seems reasonable and indeed it is generally consistent with an item interference account of OI and with data showing a release from OI when the category of stimulus changes during a recognition memory test (Criss et al., submitted; Malmberg et al., 2012). However, existing data temper our enthusiasm for this explanation. For example, Malmberg et al. (2012) included a condition where the stimulus category changed every five trials. However, this was not sufficient to observe buildup and release from OI. Further, Annis et al. (2013) included semantic tasks involving stimuli that were similar (words) and dissimilar (faces) to the stimuli in the episodic task and found no differences related to stimulus type, suggesting that similarity of the search set is not a sufficient explanation. Evidence against all three of these less interesting explanations comes from unpublished data from this project. We measured accuracy across successive questions for in-class exams over the course of a semester and found no evidence of OI in the exams. The questions from the exam were related in that they addressed similar content (e.g., one test covered short- and long-term memory chapters). Of course, the limitations of such a study prevent us from drawing strong conclusions, but the absence of OI speaks against the potential explanations of difficulty and similarity.

We therefore suggest that the best explanation of the data is the differential reliance on episodic and semantic memory to correctly answer questions. Episodic memory may be more susceptible to the sort of interference generated during testing (e.g., gradual changes in context, encoding/updating of items), whereas semantic memory is less so. Interestingly, attempting to reinstate encoding context, as suggested by many textbooks based on classic memory research (e.g., S. M. Smith, 1979) as a useful strategy for test taking, may be detrimental for tests intended to measure knowledge. That is, focusing on episodic retrieval using context information allows for interference from other information that matches the same context.

In summary, we found that successive searches of knowledge do not suffer from OI using a set of general knowledge questions. However, OI was robust when participants presumably completed the test of knowledge by relying on episodic information, namely corrective feedback, provided during Test 1. These data provide potential constraints on the relationship between episodic and semantic memory.

Author notes The authors would like to thank Shantai Peckoo for her assistance in developing the stimuli and collecting data, and Michael L. Kalish for his valuable input on the analyses and their exposition. This material is partially based upon work supported by the National Science Foundation [grant number 0951612] awarded to A. H. Criss. Data and analyses are available for download at http://sites.google.com/site/williamaue/

References

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1365–1376. doi:10.1037/a0032188

Blaxton, T. A., & Neely, J. H. (1983). Inhibition from semantically related primes: Evidence of a category-specific inhibition. *Memory & Cognition, 11*(5), 500–510.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of restricted associative responses. *Journal of General Psychology, 30,* 149–165.

Bousfield, W. A., Sedgewick, C. H. W., & Cohen, N. H. (1954). Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology, 67,* 111–118.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Core Team, R. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64*(4), 316–326. doi:10.1016/j.jml.2011.02.003

[3] We thank an anonymous reviewer for this suggestion.

Delaney, P. F., Verkoeijen, P. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). San Diego, CA: Elsevier Academic Press. doi:10.1016/S0079-7421(10)53003-2

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society, 16*(5), 748–753. doi:10.1017/S1355617710000676

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431–440. doi:10.1037/a0027373

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*(1), 85–98. doi:10.1006/jmla.2001.2798

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*(3), 306–340. doi:10.1037/0096-3445.110.3.306

Jeffreys, H. (1961). *The theory of probability*. Oxford, UK: Oxford University Press.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation, 61,* 237–284.

Karpicke, J. D., & Roediger, H. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. doi:10.1126/science.1152408

Koop, G. J., Criss, A. H., & Malmberg, K. J. (2015). The role of mnemonic processes in pure-target and pure-foil recognition memory. *Psychonomic Bulletin & Review, 22*(2), 509–516. doi:10.3758/s13423-014-0703-5

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. San Diego, CA: Elsevier Academic Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23*(2), 115–119.

Malmberg, K. J., Lehman, M., Annis, J., Criss, A. H., & Shiffrin, R. M. (2014). Consequences of testing memory. *Psychology of Learning and Motivation, 61,* 285–313.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419–457. doi:10.1037/0033-295X.102.3.419

Mueller, S. T., & Shiffrin, R. M. (2006). *REM II: A model of the developmental co-evolution of episodic memory and semantic knowledge.* Paper presented at the International Conference on Learning and Development (ICDL), Bloomington, IN.

Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage, and retrieval of item information. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 145–194). Hillsdale, NJ: Erlbaum.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19*(3), 338–368. doi:10.1016/S0022-5371(80)90266-2

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review, 120*(2), 356.

Nyberg, L., & Tulving, E. (1996). Classifying human long-term memory: Evidence from converging dissociations. *European Journal of Cognitive Psychology, 8*(2), 163–183. doi:10.1080/095414496383130

Plummer, M. (2013). rjags: Bayesian graphical models using MCMC (R package version 3-10). Retrieved from http://CRAN.R-project.org/package=rjags

Prince, S. E., Tsukiura, T., & Cabeza, R. (2007). Distinguishing the neural correlates of episodic memory encoding and semantic memory retrieval. *Psychological Science, 18*(2), 144–151. doi:10.1111/j.1467-9280.2007.01864.x

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM—A theory of probabilistic search of associative memory. In G. H. Sower (Ed.), *The psychology, of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207–262). New York, NY: Academic Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Order effects in recall. In J. B. Long & A. D. Baddeley (Eds.), *Attention and performance IX* (pp. 403–415). Hillsdale, NJ: Erlbaum.

Ratcliff, R., & Hockley, W. E. (1980). Repeated negatives in item recognition: Nonmonotonic lag functions. *Attention and performance VIII.* Hillsdale, NJ: Erlbaum.

Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review, 104*(2), 319–343. doi:10.1037/0033-295X.104.2.319

Reder, L. M., Park, H., & Kieffaber, P. D. (2009). Memory systems do not divide on consciousness: Reinterpreting memory in terms of activation and binding. *Psychological Bulletin, 135*(1), 23–49. doi:10.1037/a0013974

Roediger, H. L., III. (1973). Inhibition in recall from cueing with recall targets. *Journal of Verbal Learning and Verbal Behavior, 12,* 644–657.

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition, 2*(2), 261–269. doi:10.3758/BF03208993

Roediger, H., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning and Memory, 6*(1), 91–105. doi:10.1037/0278-7393.6.1.91

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition, 22,* 511–524.

Rouder, J. N., Speckman, P. L., Dongchu, S., Morey, R. D., & Iverson, G. (2009). Bayesian *t* test for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review, 108*(1), 257–272. doi:10.1037/0033-295X.108.1.257

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166.

Slamecka, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology, 76*(4, Pt.1), 504–513. doi:10.1037/h0025695

Smith, A. D. (1973). Input order and output interference in organized recall. *Journal of Experimental Psychology, 100*(1), 147–150. doi:10.1037/h0035513

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*(5), 460–471. doi:10.1037/0278-7393.5.5.460

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods, 45*(4), 1115–1143. doi:10.3758/s13428-012-0307-9

Tulving, E. (1972). *Organization of memory*. Oxford, England: Academic Press.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60,* 158–189.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 2*(5/6), 440–445. doi:10.1016/S0022-5371(63)80045-6