Mechanisms of output interference in cued recall

Jack H. Wilson[1] · David Kellen[1] · Amy H. Criss[1]

Abstract

The primary aim of this paper is to elucidate the mechanisms governing output interference in cued recall. Output interference describes the phenomenon where accuracy decrease over the course of an episodic memory test. Output inference in cued recall takes the form of a decrease in correct and intrusion responses and an increase in failures to response across the test. This pattern can only be accounted for by a model with two complementary mechanisms: learning during retrieval and a response filter that prevents repeated recall of the same item. We investigate how a retrieval filter might operate by manipulating the similarity of words. The data are consistent with a retrieval filter that does not operate by a global match of a potential target to previously recalled items. Results are discussed within the search of associative memory theory.

Keywords Cued recall · Output interference · Memory models · Memory · Recall

One tenet of cognition is that the capacity to retrieve experienced events can be impeded by memory for other events. Much research is devoted to studying how the encoding of memory traces interferes with the ability to retrieve a target event. Here, we focus on interference caused by retrieval itself, a phenomenon known as *output interference*. We use the term output interference to describe the general empirical observation that performance declines following episodic memory testing. The nature of that decline and the specific underlying mechanisms depend on the specific testing paradigm. For example, in recognition memory, participants discriminate previously studied targets from unstudied foils. In that task, accuracy (percentage correct in forced choice, or discrimination in single item recognition) decreases across test trials (Aue, Criss, & Prince, 2015; Criss, Malmberg, & Shiffrin, 2011; Murdock & Anderson, 1975). This pattern holds under a wide variety of circumstances including when study–test lag is not confounded with test trial in recognition and recall (Criss, Malmberg, et al., 2011; Smith, 1971). A more detailed evaluation shows that the ability to identify a target (hit rate) decreases, but the ability to reject a foil is usually unchanged

(false-alarm rate) across test trial. In free recall, output interference manifests as an increase of interresponse times over the course of test (Murdock & Okada, 1970; Patterson, Meltzer, & Mandler, 1971; Rohrer & Wixted, 1994). This suggests that different mechanisms are responsible for the patterns of output interference in recall and recognition. Whereas output interference in free recall seems to depend on the successful retrieval of an item from memory (see Tulving & Arbuckle, 1966), output interference in recognition seems to depend on the presentation of a test cue (see Criss, Malmberg, et al., 2011).

In this paper, we focus on paired associates cued recall (*cued recall*, for short), a task that has elements of both recognition and recall. In a typical cued recall task, a participant studies a list of pairs. Later, the participant is given one member of a studied pair (we refer to this item as the *cue*) and asked to generate the other member of the pair (we refer to this item as the *target*). A participant may give no response (a *response failure*) or respond with the correct item or with an incorrect item (an *intrusion*). The empirical pattern of output interference in cued recall has so far received little attention. Tulving and Arbuckle (1963) first noted output interference for short lists of noun–number pairs. They reported a decline in both correct responses and intrusions. Since then, there have been only a few reports documenting a decrease in correct responses in cued recall (including Raaijmakers & Shiffrin, 1981b, and Roediger & Schmidt, 1980) and no reports that include

✉ Jack H. Wilson
  jhwilson99@gmail.com

✉ Amy H. Criss
  amy.criss@gmail.com

[1] Syracuse University, Syracuse, NY, USA

detailed analysis of intrusions and response failures as a function of output position in cued recall.[1] It is perhaps possible that this pattern of output interference simply reflects outdated methodology where participants responded with pen and paper and could review and edit as the test proceeded. Alternatively, interference in working memory, not long-term memory, may drive the pattern of intrusions and response failures in the short-list experiments (Tulving & Arbuckle, 1966). Therefore, one goal of this paper is to establish the empirical pattern of responses across test position in cued recall.

A second goal of this paper is to better understand the mechanisms that underlie output interference in cued recall. To do so, we begin with mechanisms used to account for output interference in recognition and free recall. We consider the mechanisms postulated by the search of associative memory (SAM; Raaijmakers & Shiffrin, 1981a, 1981b) theory, an elaboration of the modal model (Atkinson & Shiffrin, 1968). In SAM, context and items are associated with one another, and the strength of the association depends on a number of factors (e.g., semantic similarity, co-occurrence, control processes). A cue or cues are used to search memory. Memory traces are sampled in relation to the strength between the trace and the cues. Search continues until a trace is recovered—that is, the details are extracted and output from memory or until the model fails to successfully sample and recover some number of times. SAM relied on two mechanisms to account for output interference—learning during retrieval and retrieval filtering.

Learning during retrieval is the idea that retrieval is an active process that results in the encoding of additional information. In SAM this process is called *incrementing* and is implemented as a strengthening of the associations between the test cue, the context of the experiment, and the recalled target. In the model, incrementing only occurs for successful recovery meaning that there is no learning when retrieval fails. In typical experiments, a single context cue is used throughout test, therefore incrementing increases the probability of sampling an item that was already recalled. Learning during retrieval was a crucial component of the SAM and played a pivotal role in accounting for part-list cueing and list length effects in free recall (Raaijmakers & Shiffrin, 1981a, 1981b).

The idea of learning during retrieval was later incorporated into the retrieving effectively from memory (REM; Shiffrin & Steyvers, 1997) framework. This allows REM to account for many of the constituent patterns of output interference in recognition (e.g., Aue et al., 2015; Criss, Malmberg, et al., 2011; Kılıç, Criss, Malmberg, & Shiffrin, 2017; Koop, Criss, &

Malmberg, 2015). However, the implementations found in SAM and REM differ. In REM, learning during retrieval occurs by updating a matching memory trace when the cue is judged to be old and by storing a new trace when the cue is judged to be new. In sum, learning during retrieval is a fundamental and well-established property of memory, as evident in empirical findings like the testing effect (e.g., Roediger & Karpicke, 2006) and models like REM and SAM, as well as common sense. After all, the world is not conveniently divided into study and test blocks. Learning and retrieval are seamlessly interwoven.

A *retrieval filter* is a control process whereby attempted retrieval of an item that has already recalled item is considered a failure. That is, if a memory trace has already been recovered and is sampled again, then that retrieval attempt is considered by SAM to be a failure, increasing the probability that memory search will be terminated. This censoring process is described in the original SAM papers (e.g., Raaijmakers & Shiffrin, 1981b, p. 412; see Fig. 4 and corresponding text in Raaijmakers & Shiffrin, 1980) and was implemented by assumption and has received little attention. We introduce the term "retrieval filter" for this mechanism. The retrieval filter was presumably implemented to prevent the system from getting "stuck" and offers a measure of how thorough the search process has been. If the search is yielding only items that it has already retrieved, then there is reason to believe that there is little new information to be found. It turns out that the retrieval filter is critical to the predicted pattern of output interference, as we demonstrate next, because it increases the occurrence of response failures across test trial. To fulfill our goal of understanding the mechanisms that underlie output interference in cued recall, we must understand how this filter operates, and that will be a major contribution of this paper.

In Fig. 1 we present simulations from SAM, implemented as described by Raaijmakers and Shiffrin (1981b) using their cued-recall parameters (see values in Fig. 1). For the simulations, we set list length of the study and test list to 20 pairs and study time at 4 s. We ran the model for 100 experiments of 80 simulated participants each. We plot bins of five test trials each, and show the average response rate and 90% credibility intervals.

The standard model implements the model exactly as described and therefore includes the retrieval filter and learning during retrieval. The standard model, shown in the left panel of Fig. 1, predicts an increase in response failures along with a decrease in correct and intrusion responses as a function of test trial. Learning during retrieval leads to stronger binding between the experimental context, cue word, and recalled word, making prior cues and prior responses more likely to be sampled again on subsequent test trials. However, if they are

---

[1] Such analyses have been conducted for category cued recall where participants study multiple single items from multiple different categories and recall items in response to a category cue (e.g., Roediger, 1973; Roediger & Schmidt, 1980; Smith, 1971). We will consider how this informs theory in the discussion.
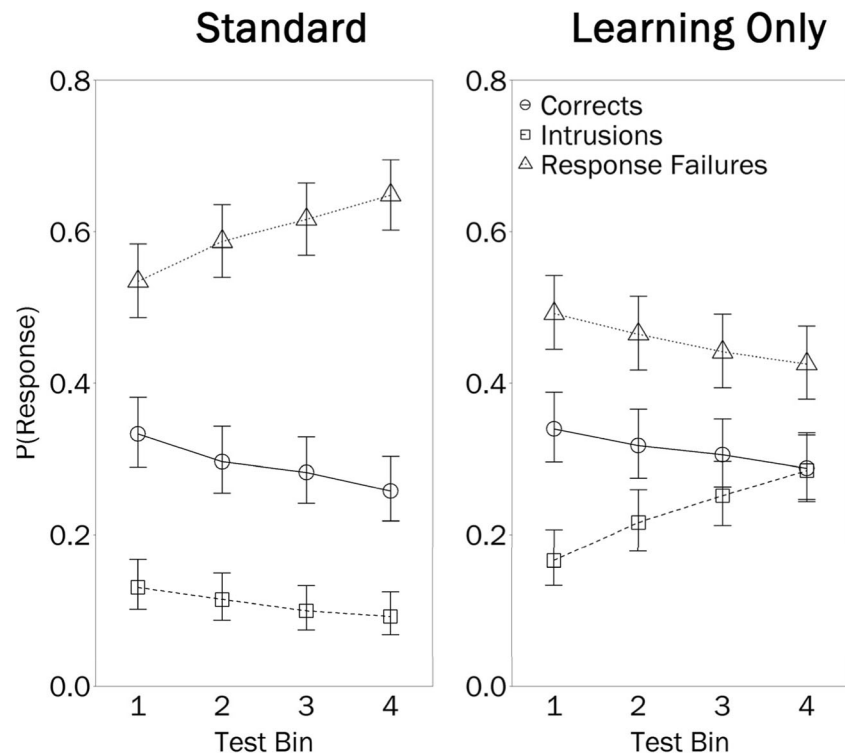
Fig. 1 SAM model predictions. The standard model includes both the retrieval filter and learning during retrieval. In the learning-only model, the retrieval filter is disabled. Parameters replicate Raaijmakers and Shiffrin (1981b). Learning rates at study: context-to-item = 0.16, interitem = 0.58. Background interitem association = 0.029, incrementing of context-to-item at test = 0.9. Study time = 4. Buffer size = 1 pair of 2 items. Retrieval attempts = 1 per test trial. Self-strength and interitem incrementing are not parameterized as they are not relevant to the cued recall simulations

sampled, they are not recalled—instead, the outcome of sampling a previously recalled item is a response failure via the retrieval filter. This results in an increase in response failures and a decrease in intrusions over test trial.

We eliminate the retrieval filter in simulations labeled *learning only*. Predictions of this model are shown in the right panel of Fig. 1. In this model, intrusions increase and response failures decrease across test trial, the opposite of what is predicted. This reflects learning during retrieval—the strength between the context, the cue word, and the recalled word is incremented upon recall. On subsequent trials, context is a cue along with a different cue word. The increased strength of association between context and those previously recalled words make them more likely to be sampled. If so, and if recovery is successful, the response is now an intrusion. Thus, each successful recall increases the probability that the recalled item or the cue word will later be output and typically those words are incorrect. To our knowledge, there are no publications showing an increase in intrusions across test block in cued recall, consistent with the standard model. However, as mentioned earlier, this could be an artifact of participants editing written responses throughout the test.

The goal of this manuscript is to document the empirical pattern of and underlying mechanisms that explain output interference in cued recall. Learning during retrieval is a well-understood mechanism, but the retrieval filter is not. Therefore, much of the focus of this paper is dedicated to revealing that process. First, we conduct a secondary data analysis evaluating the detailed pattern of output inference in cued recall. Both models predict a decrease in correct responses with test bin, but differ in their predictions of response failures (increasing for the standard model and decreasing for the learning-only model) and intrusions (decreasing for the standard model and increasing for the learning-only model). Therefore, the focus on our analysis of the empirical pattern of output interference will be on response failures and intrusions. We then conduct experiments designed to elucidate how the retrieval filter operates by manipulating the similarity of the cues or targets.

Analysis of output interference in cued recall

Wilson and Criss (2017) reported cued-recall data from five experiments with a total of 229 participants (data available at https://osf.io/4xmsx/). All experiments implemented a list-strength effect paradigm, meaning that pairs varied in the number of presentations (one, two, or four, depending on the experiment) and some study–test blocks contained a mix of strong and weak pairs, while others contained only strong or

only weak pairs. Participants studied word pairs with no distinction being made between what would later become a cue or target word. Each experiment also included perturbations of several other parameters, including study task (in some cases, no task; in some cases, forming an interactive image; in some cases, rating the difficulty of forming an interactive image), study time (1.25 s or 3 s), and list length. Some of the experiments included additional study lists that were tested with recognition or free recall. The analyses in this manuscript only consider cued recall tests that are 16 or 24 trials in length (this includes all of Experiments 1–4 and a subset of Experiment 5). In cued recall, one of the words was randomly chosen to serve as a cue and presented on the screen. Participants were asked to type the word it was studied with at their own pace. Every pair was tested (16 or 24 test trials in total, depending on the experiment).

In Cox, Hemmer, Aue, and Criss (2018), 462 participants each completed 15 study–test blocks, three blocks of each of five different test tasks including cued recall, free recall, single-item recognition, associative recognition, and lexical decision (data available at https://osf.io/dd8kp/). The study conditions were identical for all tests (excluding lexical decision, which has no study phase). Participants studied a list of 20 word pairs for 2 seconds and rated the degree of association between the words. Following study, participants completed a 45-s distractor task. The type of test was randomly ordered and postcued. This design eliminates the possibility that participants could use different encoding strategies for different tests. We analyzed only the cued-recall tests here—participants were presented with one member of each pair, randomly chosen and randomly ordered, and asked to type the item with which it was studied.

Method

Traditional analysis techniques for output interference (e.g., regression) assume linear effects. We take a more general approach and use order-constrained methods, which makes no assumptions about functional form. Order constrained inference is perfectly suited for ordinal predictions on categorical data, as in the case of output interference (for overviews, see Davis-Stober, 2009; Klugkist, Kato, & Hoijtink, 2005; Silvapulle & Sen, 2011). For example, one hypothesis we will evaluate is whether each successive test bin has more response failures than the last. Likewise, we will evaluate whether each successive test bin has fewer intrusions than the last. These specific hypotheses are compared with completely unconstrained orderings of conditions. Details about order constrained inference applied to this design follow.

The proportions of response failures (RF), correct responses (C), and intrusions (I), associated with each test bin n, with $n = 1, \ldots N$, are assumed to follow a multinomial distribution with probability parameter vector $\boldsymbol{p_n} = \left[p_n^{(RF)}, p_n^{(C)}, p_n^{(I)} \right]$, with $p_n^{(RF)} + p_n^{(C)} + p_n^{(I)} = 1$.

Based on the simulations reported in Fig. 1, we are able to establish a set of ordinal predictions at the level of the probability parameters of each response category. The standard model hypothesizes that any for any two response bins n and $n + 1$:[2]

$$p_n^{(C)} \geq p_{n+1}^{(C)},$$
$$p_n^{(I)} \geq p_{n+1}^{(I)}, \text{and}$$
$$p_n^{(RF)} \leq p_{n+1}^{(RF)}.$$

These inequalities capture the notion that both correct responses and intrusions decrease across test bins, whereas response failures increase.

The learning-only model yields a somewhat different hypothesis—that correct responses and response failures decrease across test bins and intrusions increase.

$$p_n^{(C)} \geq p_{n+1}^{(C)},$$
$$p_n^{(I)} \leq p_{n+1}^{(I)}, \text{and}$$
$$p_n^{(RF)} \geq p_{n+1}^{(RF)}.$$

Hypothesis testing was conducted under a Bayesian framework (for reviews, see Gelman et al., 2014; Lee & Wagenmakers, 2013). Relative evidential support for a given hypothesis against a competitor can be quantified by means of Bayes factors (Kass & Raftery, 1995). In the present case, the data follow a joint multinomial distribution. A reasonable choice for an uninformative prior for each three-outcome multinomial is the Dirichlet distribution with concentration parameter vector $\boldsymbol{\alpha} = [1, 1, 1]$. The Dirichlet distribution is a conjugate prior, which means that when updated with multinomial-distributed data, the resulting posterior is also a Dirichlet distribution with known concentration parameters. We compute evidence for the learning-only model and the standard model (H_i and H_j) each compared with a vacuous alternative hypothesis (H_s) in which all probabilities are completely unconstrained. As shown by Klugkist et al. (2005), the Bayes factor quantifying the evidential support for the H_i relative to H_s can be directly computed from the proportion of samples from the posterior and prior distributions that conform to the constraints established by H_i:

$$BF_{i,s} = \frac{P(posterior \in H_i)}{P(prior \in H_i)}.$$

---

[2] The reliance on inequality constraints has two important advantages: First, it releases us from having to assume a specific functional form (e.g., increase is linear) that is not part of the theoretical predictions. Second, the probabilities permitted by the inequality constraints form a convex polytope. This means that, unlike with other approaches (e.g., Estes, 1956), the hypothesis being tested cannot be spuriously rejected because of the use of aggregate data (for a review, see Regenwetter & Robinson, 2017).

Also, the Bayes factor for any two hypotheses of interest can be directly computed from the Bayes factors comparing them against H_s; $BF_{i,j} = BF_{i,s}/BF_{j,s}$. Bayes factors indicate the ratio of evidence in favor of a hypothesis, more extreme values indicate stronger support.

Results and discussion

The data from Wilson and Criss (2017) and Cox et al. (2018) are illustrated in Fig. 2. Visual inspection shows that the ability of the standard-model predictions to capture the observed response outcomes was virtually perfect. In contrast, the data were completely at odds with the predictions of the learning-only model. These results are captured by the Bayes factors obtained with the data from Wilson and Criss (2017) (LL = 16: BF = 1,889; LL = 24: BF = 202) and Cox et al. (2018) (BF = 309), indicating strong support for the standard model relative to the learning-only model. The difference in Bayes factors between the two studies can be attributed to their difference in subject-sample sizes.

Does the retrieval filter fail?

To help understand how a retrieval filter might work, we investigate how often participants recall words that had already been recalled. Recalling an item more than once in a given test list indicates that the retrieval filter failed. We simply counted the proportion of unique recalls for each participant on each test list. For the purposes of this analysis, a word was counted

as a repetition only if it exactly matched another output. In Wilson and Criss (2017), 5% of all responses were duplicates of a word recalled earlier on that test list, and the remaining 95% of responses were unique words. In Cox et al. (2018), 3% of responses were repeats of a word recalled earlier in the list. In summary, the data show is that a filter is necessary for SAM to account for the specific pattern of output interference in cued recall. Now, let us turn to our primary focus, which is to help illuminate how a retrieval filter might operate.

Overview of Experiments 1–3

Having established the necessity of a retrieval filter, at least within the SAM framework, we now ask how such filter might work. There is no mechanism for the filter proposed in SAM. We consider the two viable options borrowed from mechanistic principles in the memory modeling literature. One possibility is that a tag, feature, or contextual element is added to a memory trace upon recall. The retrieval filter might operate by evaluating that local information within a trace. Another possibility is that the retrieval filter relies on a global matching process (Humphreys, Pike, Bain, & Tehan, 1989). For example, the retrieval filter might operate by comparing the to-be-retrieved item to the list of all traces outputted during a test, and if the to-be-retrieved item matches beyond a threshold, then the response is censored. A global matching process approach would predict that, for instance, outputting a list of similar responses (e.g., targets of the same category) would
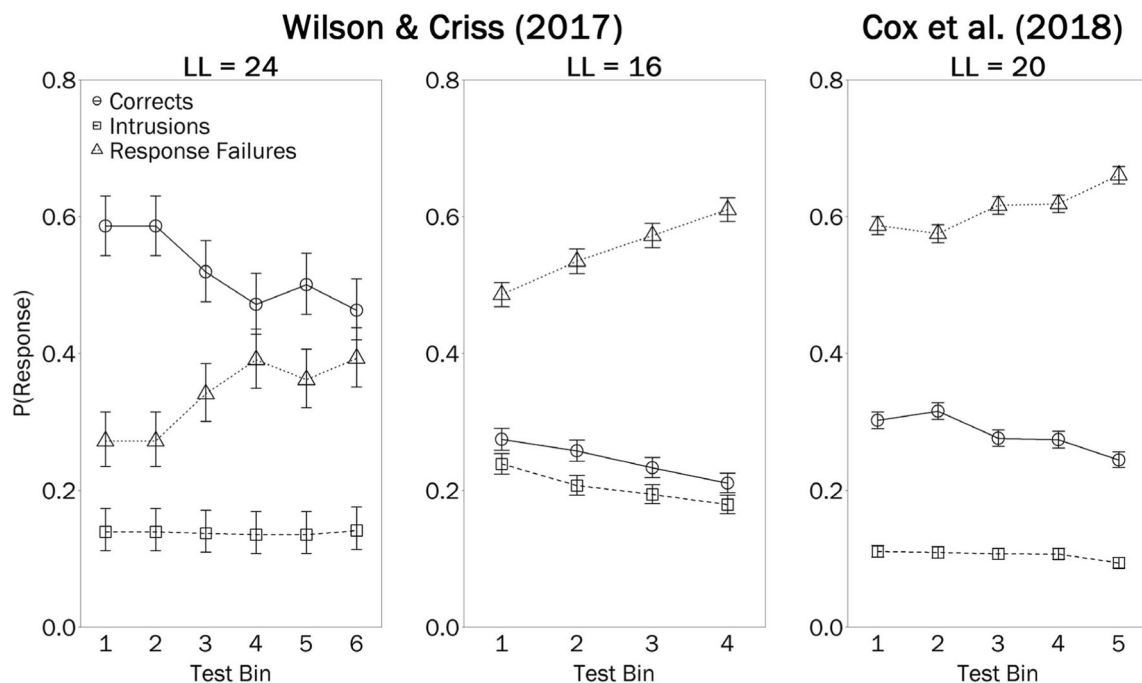

Fig. 2 Mean response proportions (circles: correct responses; squares: intrusions; white triangles: response failures; black triangles: output-bound accuracy [OBA]) and their respective 95% credibility intervals from the data reported by Wilson and Criss (2017; left side) and Cox et al. (2018; right side)

result in a greater global match, which in turn would yield a decrease in responses. As a concrete example, suppose the participant responded with a series of six animals. On the seventh trial, a candidate to-be-retrieved item is compared with the already recalled items. If that candidate is also an animal, then the global match will be high, exceeding the threshold. The memory system concludes (erroneously in this case) that the candidate has already been recalled, and therefore considers this attempt a failure to retrieve. This same mechanism would also predict that changing the target category (e.g., if the candidate for retrieval on the seventh trial was a tool rather than an animal) would result in a release from output interference. A filter that operates by a local tag would make neither of these predictions.

A release from output interference design allows us to test these two potential mechanisms for a filter. In a typical design (Criss, Saolmão, Malmberg, Aue, & Claridge, 2018; Malmberg, Criss, Gangwani, & Shiffrin, 2012), participants study a list of words from two different categories and are then tested on that list. Performance decreases with test trial. However, if the words from one category are tested in succession, then switching to a different category results in a release from output interference. Release from output interference is then observed in the form of an increase in performance when the second category begins to be tested. In other words, the interference that causes accuracy to decline over test position is released when the category switches and performance returns to the level that it was on the first test trial.

Release from proactive interference paradigms (e.g., Gardiner, Craik, & Birtwistle, 1972; O'Neill, Sutcliffe, & Tulving, 1976; Wickens, Born, & Allen, 1963) are similar to release from output interference paradigms. The key difference is that, in release from proactive interference designs, participants are repeatedly studied and tested on the same kind of material over multiple study–test blocks, with the switch occurring after a study–test block. In release from output interference designs, the study conditions are identical, and the switch is in the middle of the test. Therefore, the mechanism behind release from output interference is necessarily limited to retrieval.

An important aspect of cued recall is that the test cue and correct response are different words and therefore can be subject to different manipulations. For instance, we can categorize either the cues or the targets. A retrieval filter operating on a global match makes two predictions. First, when targets are categorized and tested in a blocked fashion, a series of similar items are recalled successively. This results in a buildup of the global match over output position. Targets that should be recalled later in the test are more likely to be rejected by the filter because of the increasing global match. The outcome is an increase in response failures compared with a situation where targets are not categorized. Second, this interference caused by successive testing of related items should be

released when the category of the target responses is switched during test.

General method

Participants

All subjects were undergraduates at Syracuse University participating for class credit. Sample size was chosen to be roughly double the sample size in prior studies evaluating output interference in cued recall.

Materials

Stimuli consisted of a subset of the Battig and Montague (1969) categories updated by Van Overschelde, Rawson, and Dunlosky (2004). The word pool consisted of the 13 most common unique words from each of 28 categories. We excluded categories that shared words or consisted primarily of low-frequency words.

Design

Prior to study, participants were informed they would be tested on their memory for the word pairs, but were given no information about the nature of the stimuli, information as to what words would later serve as cues or targets, or structure of the test list. Figure 3 shows a schematic of the design. This was a 2 × 2 mixed design. Participants were randomly assigned to one of two groups. For one group, the categorized words served as the test cues (*cues categorized*), and for the other group, the categorized words served as the to-be-retrieved targets (*targets categorized*). All participants received one test list that was *blocked*, such that all the word pairs of one category were tested before all the pairs of the other category. The other list was presented in a *mixed* fashion, with pairs from categories intermixed, and order of blocked versus mixed testing was randomly selected for each participant.

During the study phase, participants were instructed to study each pair by forming an image containing the pair. The 24 study pairs were constructed by randomly selecting 12 items from each of two categories and pairing them with one item chosen at random from each of the remaining categories. Pairs were presented side by side centered on the screen, for 2.50 s with an interstimulus interval of 0.25 s. This is a common procedure to ensure reasonable accuracy (e.g., Criss, Aue, & Smith, 2011; Hockley & Cristi, 1996a, 1996b; Wilson & Criss, 2017). The left–right order of the items and the study position were randomly assigned. A 60-second distractor math task separated study from test.

On each test trial, participants were probed with one member of each pair and prompted to generate the other member of the pair or indicate that they did not know the response. All 24
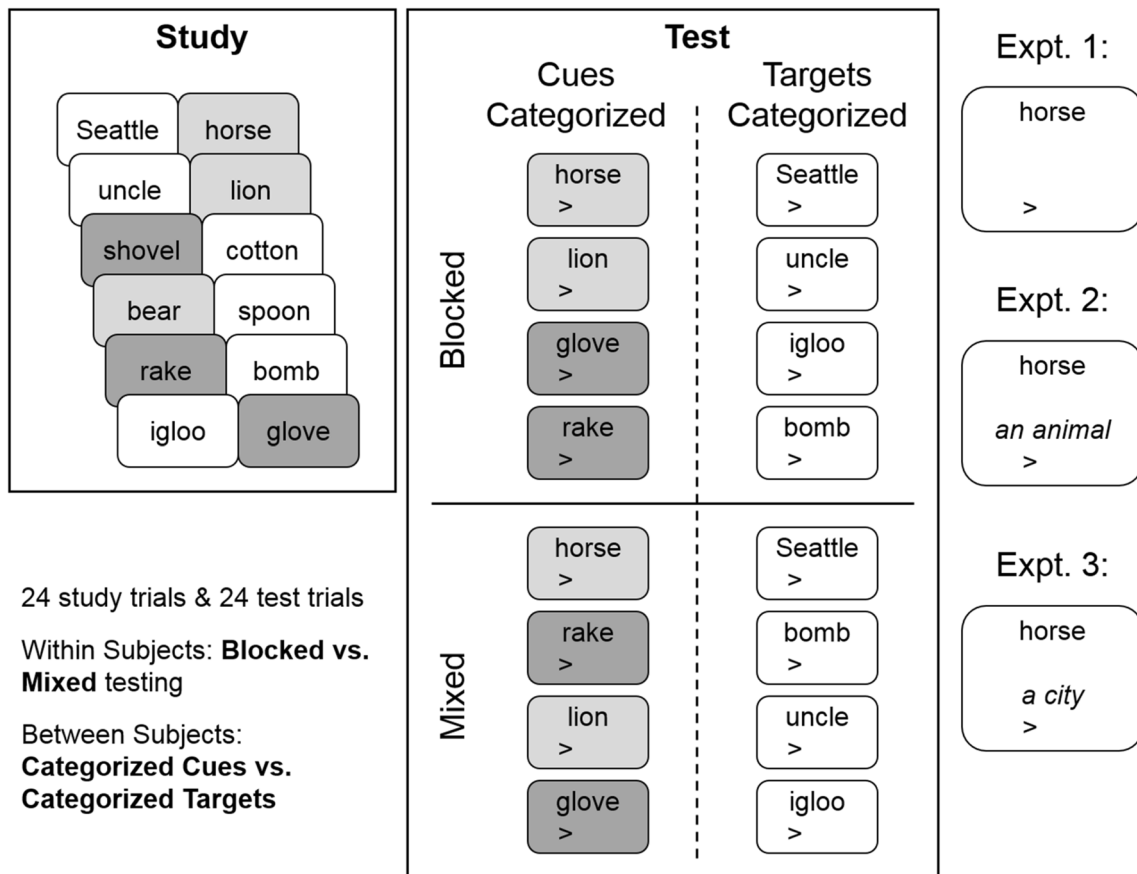
Fig. 3 Method schematic for Experiments 3 through 5. Darkly shaded words are of one category (tools), lightly shaded words are of another (four-legged animals), and unshaded words are randomly sampled from the remaining categories. This shading is for the benefit of the reader. In the experiment, all words were presented with a white background. The ">" symbol indicated to participants where their typed response would be shown

pairs were tested. A cue word was presented with a prompt to type (>) beneath it. The answer appeared immediately to the right of the prompt. Participants were instructed to type "idk" if they did not know the answer. Participants were given unlimited time to type their answer. When they finished each response, they pressed the return/enter key, and the screen went blank for 0.10 s and before the next cue word appeared.

Analysis

Responses were classified into one of three types: response failures for test trials, where no word was recalled; correct responses, where the correct target (allowing for misspellings, changes of tense, and changes of plurality) was recalled; and intrusions, where participants recalled something that was not the target. Intrusions included both intralist (i.e., presented on the list but with a different cue) and extralist (i.e., not presented on the list) responses. A characterization of the intrusions by type is presented in the Appendix.

We will once again rely on order-constrained methods. However, the testing of release in output interference raises issues that encourage us to use a convenient parameterization

of the multinomial distributions. The issue is that because the probabilities in vector $\boldsymbol{p_n} = \left[p_n^{(RF)}, p_n^{(C)}, p_n^{(I)}\right]$ sum to 1, they are not independent. This means that one cannot independently test the effects of release on each of these probabilities. One solution is to reparameterize parameter vector $\boldsymbol{p_n}$ such that we consider only two independent probabilities: $\theta_n^{(RF)}$, the probability of a response failure, and $\theta_n^{(C)}$, the probability of a correct response, *conditional* on the fact that *some response was produced* (i.e., there was no failure). This conditional probability provides us with an estimate of response accuracy that is functionally independent from response failure. Note that the probability of an intrusion, conditional on the fact that a response was given, is simply $1-\theta_n^{(C)}$. Formally:

$$p_n^{(RF)} = \theta_n^{(RF)},$$
$$p_n^{(C)} = \left(1-\theta_n^{(RF)}\right) \times \theta_n^{(C)},$$
$$p_n^{(I)} = \left(1-\theta_n^{(RF)}\right) \times \left(1-\theta_n^{(C)}\right).$$

To be clear, this reparameterization does not introduce difficulties in describing the data in any way whatsoever. Based on the results from the previous reanalysis, we can establish

the hypothesis of *output interference with no statement on release*, according to which there is a systematic increase of *response failures*, but nothing is said about the occurrence of a release. This is achieved by having distinct sets of inequalities for the preswitch and postswitch test bins. In our data, there are six test bins, with a switch between Bins 3 and 4, which leads us to expect:

$$\theta_1^{(RF)} \leq \theta_2^{(RF)} \leq \theta_3^{(RF)}, \text{ and}$$
$$\theta_4^{(RF)} \leq \theta_5^{(RF)} \leq \theta_6^{(RF)}.$$

Note that there is no inequality between test–trial Bins 3 and 4. In contrast, in order to test for the presence or absence of a release from output interference, we focus exactly on Bins 3 and 4. Specifically, if there is a release from output interference at the level of response failures, then one should expect a reduction in the proportion of response failures:

$$\theta_3^{(RF)} \geq \theta_4^{(RF)}.$$

The alternative hypothesis of no release corresponds to the opposite inequality:

$$\theta_3^{(RF)} \leq \theta_4^{(RF)}.$$

Let us now turn to the correct responses and intrusions. Again, note that under our reparametrization, the probability of a correct response, given that either a correct or an intrusion response was made, is captured by parameter $\theta^{(C)}$. Based on the simulations with the standard model illustrated in Fig. 1, we assume that the buildup of output interference leads to a decrease in the conditional probability of a correct response:

$$\theta_1^{(C)} \geq \theta_2^{(C)} \geq \theta_3^{(C)} \text{ and}$$
$$\theta_4^{(C)} \geq \theta_5^{(C)} \geq \theta_6^{(C)}.$$

Again, we there is no inequality linking Bins 3 and 4, which means that nothing is said about the occurrence of a release between the two. The hypothesis that a release took place is then captured by the inequality

$$\theta_3^{(C)} \leq \theta_4^{(C)},$$

whereas the absence of a release is captured by the opposite inequality:

$$\theta_3^{(C)} \geq \theta_4^{(C)}.$$

Experiment 1

Participants

Participating in this experiment are 104 undergraduate students at Syracuse University who participated for class credit,

equally distributed between the cues categorized and targets categorized groups.

Design

The design is described in the General Methods. The memory probe consisted of only the cue word.

Results

Standard output interference was observed here just as in earlier studies. The Bayes factors in Table 1 obtained closely follow the results illustrated in Fig. 4. When the categorized words served as test cues, we found strong evidence for output interference at the level of response failures (i.e., $\theta^{(RF)}$). Release from the increase in response failures occurs only in the cues-blocked condition. In this condition, response failures increase over test bin, then decrease at the point the cue category changes, then increase again. The corresponding decrease in responses is at the same magnitude for both correct and intrusion responses (i.e., outbound accuracy, the conditional rate of giving a correct response given that any response is made, is constant across test bin).

Discussion

Consistent with data from prior studies, output interference in cued recall is expressed as an increase in response failures. This pattern of results is consistent with the need for both learning during retrieval and a retrieval filter as included in the standard implementation of SAM. We seek to understand how such a retrieval filter might operate. To do so, we look at performance when the target items or cue items are related.

Blocking the cue word by category such that the test cue is from the same semantic category on several successive trials affects performance. Specifically, response failures decrease at the category switch point, indicating that there is a renewed ability to provide responses (correct or incorrect) when the cue changes to a different category. Subsequent trials with cues from the new category result in an increase in response failures once again, characteristic of output interference. These data are in line with a cue-dependent retrieval mechanism. When the cue changes, so too does the resulting search of memory. For example, items from category one occupy a shared semantic/episodic space, and repeated searching of that space fails to find candidates available for sampling and retrieval. When the cue changes to a different semantic cue, new traces are available for sampling and retrieval. This is consistent with SAM and many other global matching models, but these data do not speak to potential mechanisms for the retrieval filter.

Blocking the test list so that the correct target response is from the same semantic category on several successive trials

Table 1 Bayes factors (BFs) for release in Experiments 1 through 3

	Response failures		Conditional correct responses	
	Evidence for output interference	Evidence for release	Evidence for output interference	Evidence for release
Experiment 1				
Cues mixed	9.3	**0.06**	0.14	1.2
Cues blocked	**18**	**41**	**0.04**	0.36
Targets mixed	1.3	0.23	3.0	3.3
Targets blocked	0.95	**0.03**	0.13	0.65
Experiment 2				
Cues mixed	**13**	1.5	0.56	1.0
Cues blocked	**23**	**120**	**0.06**	**0.06**
Targets mixed	**10**	0.53	0.27	0.86
Targets blocked	**12**	**12**	3.2	**120**
Experiment 3				
Cues mixed	7.4	0.18	1.7	0.12
Cues blocked	**11**	3.1	5.9	**17**
Targets mixed	**15**	**0.01**	0.14	1.0
Targets blocked	**32**	**22000**	**0**	**0**

BFs rounded to two significant figures. BFs larger than 1.0 denote evidence for the pattern, and BFs smaller than 1.0 denote evidence against the pattern. Boldface values are larger than 10 or smaller than 0.1

has no discernable effect on performance. This is inconsistent with a filter that relies on a global match. A mechanism of that sort would suffer interference from the previously generated target responses and a corresponding release from interference when the category of the target response was changed.

Why does blocking by target response not affect performance, especially given the apparent necessity of a retrieval filter to predict output interference in cued recall? One possibility is that participants do not realize that the target items are from a shared category. In Experiment 3, we ensure that participants notice the relationship between successive targets by presenting the cue word and the category label of the target. Experiment 2 presents the complementary control condition where the cue category is presented in addition to the cue.

Experiment 2: Redundant cue

Participants

One hundred and sixty-one people participated in this experiment—81 in the cues group and 80 in the targets group.

Design

The only difference from Experiment 1 is that participants were also presented with a Battig and Montague (1969) category label for each test cue at test. For example, a participant who studied the word pair "leather–niece" would be given

both the word ("leather") and its category ("A type of fabric") as part of the test stimulus (see Fig. 3). Participants were informed immediately following the distractor task that the category of the cue would be provided as a hint.

Results

Performance in Experiment 2 (see Fig. 5) shows a similar pattern of output interference and release as Experiment 1. Once again, the Bayes factors shown in Table 1 show a build-up of output interference at the level of response failures, but this time the evidential support is found in all conditions. We also find strong evidential support for the presence of a release the blocked conditions, both when cues and targets are blocked at test. In terms of conditional accuracy, the only clear effects were found in the cues-blocked condition: When cues were blocked, conditional accuracy increased across test bins and suffered a decrease between Test Bins 3 and 4.

Discussion

The results of this study replicate the general pattern of results reported in Experiment 1, suggesting that the category label for the cue was indeed redundant. Presumably, participants were either aware of the category of the cue and used that in their memory search, or the category is subsumed by the cue word itself and provides no new information for the memory search.
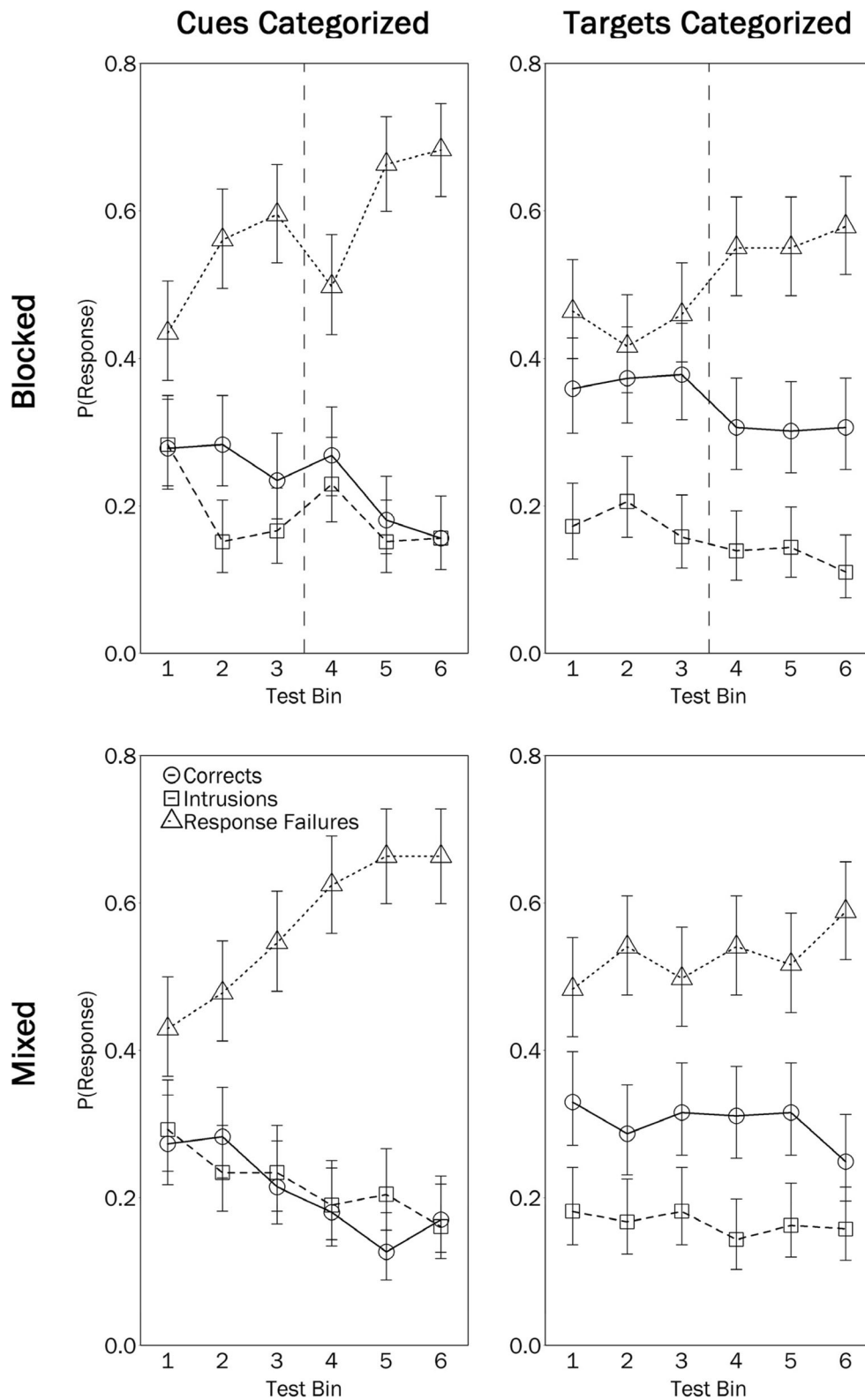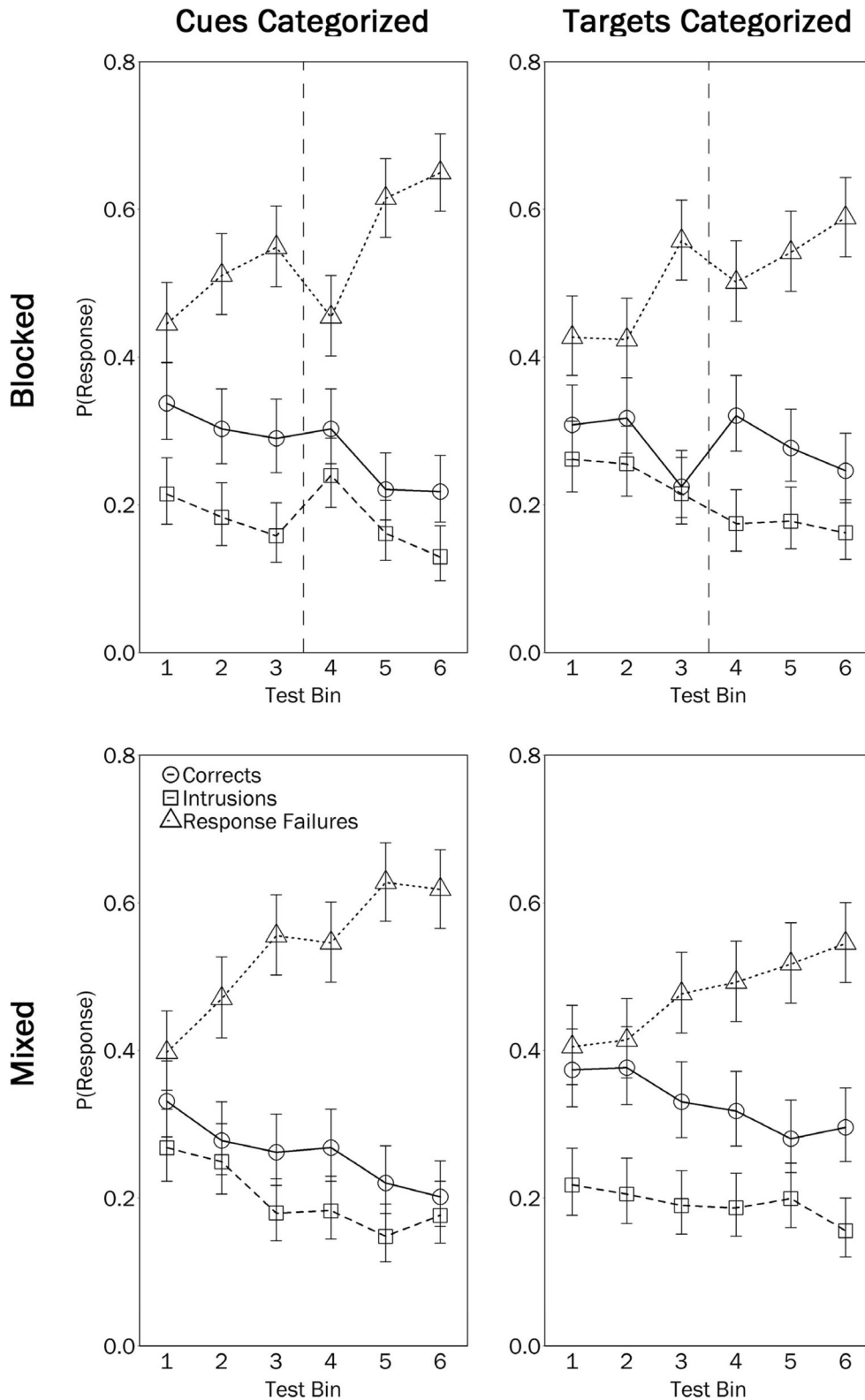
Fig. 4 Binned response proportions (circles: correct responses; squares: intrusions; triangles; response failures) and 95% credibility intervals for Experiment 1. Dotted line marks the point where category changed during blocked testing

A potentially major difference is the response failures for the targets-blocked condition. In Experiment 2, we find

evidence for a small release from output interference when the category of the targets changes. It could be that this is a

Fig. 5 Binned response proportions (circles: correct responses; squares: intrusions; triangles; response failures) and 95% credibility intervals for Experiment 2. Dotted line marks the point where category changed during blocked testing

statistical anomaly or it could be a meaningful pattern. This pattern is expected if the filter operated via a global match

process. Why this might occur when a redundant cue is provided is unclear. Perhaps there are cases when the cue is not

remembered, which would normally result in a response failure. These response failures might be avoided by searching with a general category cue rather than the specific item. These data offer mild support for a filter that operates based on global matching but we do not find the evidence convincing given that this pattern was not observed in Experiment 1.

Experiment 3: Informative cue

Participants

A total of 136 people participated in this study, with 67 in the cues group and 69 in the targets group.

Design

The only difference between this experiment and Experiment 2 is that the category information provided at test informed participants the category of the target, rather than the cue. For example, a participant who studied the word pair "leather–niece" would be given both the word "leather" and the target category "A relative" as part of the test stimulus (see Fig. 3). Participants were informed as such immediately following the distractor task.

Results

Consistent with all data reported so far, output interference was observed in all conditions in terms of an increase in response failures. As in Experiments 1 and 2, the Bayes factors in Table 1 indicate a clear evidential support for increase in response failures across test bins, consistent with the hypothesis of a buildup of output interference. We also see strong evidence for a release from output interference at the level of response failures in the blocked tests. Release from output interference in the response failures was found for the targets-blocked condition (see Fig. 6). Intrusions markedly decrease over test bin and increase at the point the target category changes in this condition. Correct responses decline over test trial in this condition do not show release at the point of the target category change. This results in output bound accuracy pattern over test bin that parallels the response failures. In the cues-blocked condition, we found no release for response failures. Both the lack of clear release for the cues-blocked condition and the isolation of release to intrusions for the targets-blocked condition are different from the patterns that we observed in Experiments 1 and 2.

Discussion

This experiment replicates the prior finding that an increase in response failures along with a decrease in correct responses across test bin. In contrast to Experiments 1 and 2, we observe a sharp decline in responses when the targets are categorized and release from the increase in response failures for categorized targets paired with release for intrusions. This is in contrast to Experiments 1 and 2, where the category of the cue, not the target, caused a small change in response rates. However, the conceptual underpinning is quite similar. Here, the cues presented to memory are a cue word and the category label for the target item. Because targets are from the same category, this label repeats for a series of successive trials (and response failures increase across these trials) before changing to the other category. At the change point, response failures fall to a much lower level. This suggests that the cue helped participants generate candidate responses, but this help waned over trial, presumably as candidate responses were exhausted.

These results are closer to the type of pattern we would expect from a filter that is based on a global match to generated responses. If the system compared a potential response with a set of already-generated responses prior to output, then we would expect to see interference that grows as the list of outputted targets from the same category grows. However, we only observe this pattern when the category of the targets is presented as a cue. This suggests that the filter is not critical, but rather that the cues used to search memory are critical to the pattern of data. In this paradigm, the cue information includes both the category of the target item and the cue word studied with the target. These cues provides multiple routes to the target item or set of potential target items (i.e., members of the category). A change in the cue changes the set of potential target items.

What makes this cue-based explanation of the data more compelling is the lack of release when cues are blocked by category. The cue word possesses the same category information, but the category of the target changes from test trial to test trial. The core of the cue-based explanation is that the nature of the search and thus the course of output interference depends on the information given by the test probe. In Experiments 1 and 2, this information consists of the cue word and the context of study and test. When the cue word is repeatedly from the same category, the search is loosely restricted to the set of 12 pair memories with the words from that category. This results in steep output interference and, further, release when the category of the cue changes. In this final
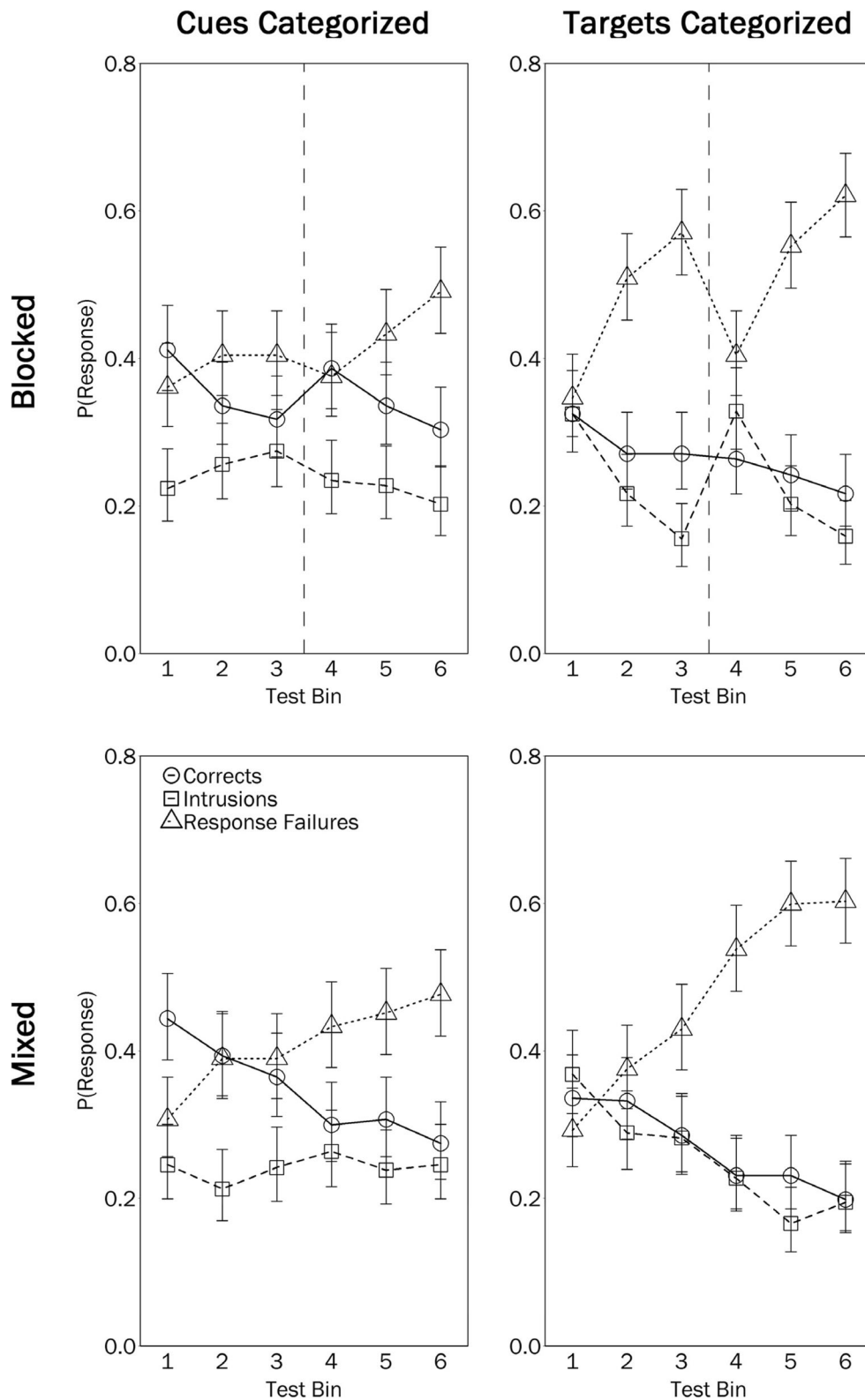
Fig. 6 Binned response proportions (circles: correct responses; squares: intrusions; triangles; response failures) and 95% credibility intervals for Experiment 2. Dotted line marks the point where category changed during blocked testing

experiment, the target category is also part of the test probe. This expands the scope of the search beyond that

set of 12 pair memories because the category of the target mismatches all the other pairs. This makes the nature of

the search more like when the cues are not categorized in the first place, as in the targets categorized group. Therefore, the pattern of responses over test trials resembles the targets categorized group.

General discussion

The primary objective of this manuscript was to better understand the mechanisms underlying output interference in cued recall. The empirical pattern of data is that response failures increase across trial for this test task. This is consistent with a model that relies on both learning during retrieval and a retrieval filter, at least within the SAM framework. The process for learning during retrieval is well described in SAM, but the retrieval filter was implemented by assumption, and no mechanism was specified.

We considered two potential mechanisms for a retrieval filter. One possibility is that the filter operates by a global match process, comparing the candidate for retrieval to a set of already retrieved responses and rejecting the candidate if the global match to the set exceeds some threshold. A second possibility is that the filter operates on a feature or tag added to the memory trace upon retrieval. The filter could check a candidate for retrieval and reject only those candidates that are marked as having been recalled.

We conducted several experiments to discriminate between these two possible mechanisms. The critical manipulation was whether or not the target items were members of the same category and whether the same-category items were blocked during retrieval or randomly intermixed with items from another category. We found that release from output interference occurs when the nature of the test stimulus, usually the *cue* word, changes systematically from one half of the test list to the other. Release from output interference generally does not occur when the category of the *responses* systematically changes from one half to the next, unless that information is part of the test cue. In all, this is consistent with output interference as a cue-driven phenomenon. We found no evidence suggesting that the retrieval filter was based on a global match to those items retrieved earlier in the test list.

On the retrieval filter

The mechanism of SAM's retrieval filter is not specified: The model simply makes recalled words unrecallable by assumption. How the memory system is aware of what has or has not been recalled is not stated. We considered two options for how the memory system makes this determination: by using either a global match to a set of recalled items or a local match to some property of a to-be-recalled item.

On the one hand, the filter could test potential responses against a global match to prior stored responses. For example, responses that were generated over course of testing are stored in long-term episodic memory. The proposed response is then checked against that list, just as in recognition memory tests. If the proposed response matches, then it is withheld; otherwise, it is output. This leads to a number of predictions. A candidate response should be subject to item noise (e.g., list length, list strength, category effects). Specifically, this means that generating several items from the same category should increase noise and decrease the accuracy of producing the correct word. Additionally, when the category of potential responses changes (as in the target categorized blocked conditions), accuracy should rebound. In contrast to these predictions, output interference and release appear largely decoupled from the to-be-recovered content unless some information about that content is included in the test stimulus.

On the other hand, the filter could be using a local match. A local match retrieval filter would use some kind of tag or property of the to-be-recalled memory to determine whether or not it had been recalled. This filter is not reliant on information from other recalled items, hence the name. A local match does not predict release from output interference contingent on the semantic relationship between recalled items. For this reason, a filter based on a local match can better account for our data. An example of a local match process is that a memory trace is tagged to indicate that it was output. The retrieval filter would then check the candidate memory to see if it had a tag, and if so, the candidate would be withheld. This is consistent with the production effect, the finding that words that are vocalized are better remembered than words that are not produced (Forrin, MacLeod, & Ozubko, 2012; MacLeod, Nigel, Hourihan, Neary, & Ozubko, 2010). Critically, the production effect is only present for mixed lists but not pure lists which is similar to the circumstances of a recall task where memory is imperfect (cf. Fawcett, 2013). Of course, if the number of recalled items is relatively small, then the items could be held in a buffer and scanned (Hacker, 1980), which might result in data like those observed here. This is an interesting possibility to pursue in future research.

Absent another compelling explanation, the retrieval filter survives testing in cued recall. This filter does not operate under the principle of global matching according to the presented data. This leaves us with the tentative conclusion that some content of the accessed to-be-retrieved item is being compared with some other criterion to decide whether the item has been previously recalled, without influence from other

recalled items. It seems reasonable that this filter is a critical component of the control process governing retrieval in cued recall.

On cue-based retrieval

Consistent with prior literature, the presented experiments demonstrate that output interference is dependent on the relationship between the cue and the contents of memory. In Experiments 1 and 2, we observe a buildup of response failures that decreases when the category of the cue word changes. In Experiment 3, we pair a cue word with a description of the category to which the target word belongs, and we see release from output interference when the category of the target word (now provided as a cue) changes. Overall, these patterns reflect how output changes in response to the cue, or, in other words, how the cue drives memory search.

In SAM, the relative degree of binding between the test stimuli—in cued recall, this is usually the cue word and the context of study–test—and the contents of the long-term store drive what is sampled and therefore what is recovered, and where in memory learning occurs. This basic structure of cue-dependent retrieval accounts for the critical patterns observed in our data. Release from output interference occurs when the test stimuli consistently points to the same set of pairs over the course of changes before changing to a new set. Release happens because consistently sampling from the same set of pairs limits learning and interference to those pairs. When the set of pairs being sampled from changes, release is observed simply because little or no learning has taken place in that new set and few or none of the items from that set have been recovered. In situations where sampling is not so restricted, release is not observed because the necessary conditions are not in place.

This account is related to findings of output interference in category-cued free recall (Roediger, 1973; Roediger & Schmidt, 1980; Smith, 1971). In this task, participants study a list of words from a number of different categories. Later, as a test trial, participants are given a category as a test stimulus and instructed to output all the words they remember studying from that category. Following completion of that trial, participants given another category, and so on. The proportion of recalled words from a category declines over test trial: categories tested first have the greatest recall percentages, and categories tested last have the fewest. The studies also show that the (rather low) rate of extralist intrusions stays flat or slightly increases over test trial.

This decline of correct responses was expected by SAM and is claimed as an early success of the model.

SAM models the task by completing a free-recall sample-recovery loop using the context of study–test and the category cue as test stimuli. Items that are members of the category have strong associations to the category cue, but items outside that category also have some residual degree of association with that cue. It is therefore possible in the model to incorrectly sample a noncategory member and the stronger the item's binding to the context the more likely this is. Learning at test therefore increases the probability of incorrectly sampling recalled members of a category during later testing of a different category. Sampling is a competitive process, and therefore the odds of sampling a category member, conversely, decline. As more and more categories are tested, more and more recalled items are strengthened, and the probability of sampling items from the category gradually declines. As for intrusions, note that these are, by task design, extralist intrusions. The set of extralist items is far larger than the set of intralist items, so interference effects are mitigated.

Summary

In all, then, output interference in cued recall is characterized generally by a decline in correct responses and intrusions and an increase in response failures. This pattern is driven by learning during retrieval and a retrieval filter that operates based upon a local match of the to-be-recovered item.

Open practices statement None of the reported experiments were pre-registered. Data are available for download at <https://osf.io/bwvqe/>

Appendix

Characterization of intrusions

We analyzed whether intrusions were other words from the studied list (intralist) or not (extralist). We observed a difference in the nature of the intrusions between experiments (see Fig. 7). In Experiments 1 and 2, intrusions did not differ much in terms of type (intralist vs. extralist) and did not differ much across condition.

However, in Experiment 3, the pattern of intrusions differed by type and condition. In Experiment 3, a cue word was provided along with the label of the category
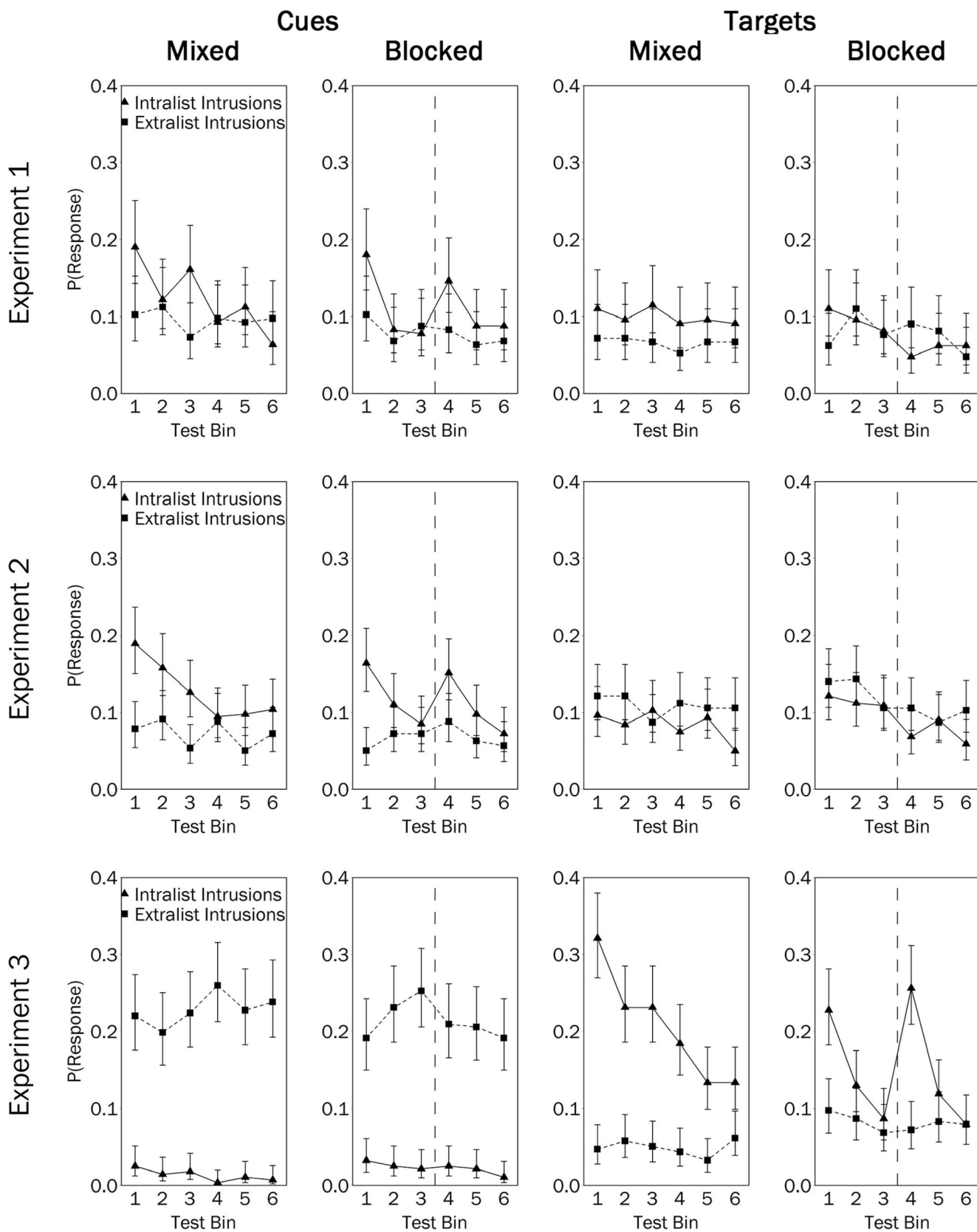
Fig. 7 Extralist (squares) versus intralist (triangles) intrusions by condition. Error bars give 95% credibility intervals. Dotted lines denote the point of change in category during blocked testing

of the target word. For example, a participant who studied the word pair "leather–niece" would be given both the word "leather" and the category of the target "A relative." When the cues shared a category (e.g., the cues were from the category "a type of fabric"), there were many more extralist intrusions than intralist intrusions. This suggests that participants were generating members of the target category in response to the memory prompt. In the example above, a participant might guess "mother." In this design, the targets did not share a category and therefore guesses of this type would all be extralist intrusions. In the condition where targets were categorized (e.g., all target items were from the category "a relative"), then guesses of these types would often be intralist intrusions. Overall, these data suggest that the cues directed the memory search to a relevant subset of items.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation: Advances in Research and Theory, 2,* 89–195. https://doi.org/10.1016/S0079-7421(08)60422-3

Aue, W. R., Criss, A. H., & Prince, M. A. (2015). Dynamic memory searches: Selective output interference for the memory of facts. *Psychonomic Bulletin & Review, 22*(6), 1798–1806.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 80,* 1–46.

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General, 147*(4), 545–590.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language, 64,* 119–132.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316–326.

Criss, A. H., Salomão, C., Malmberg, K. J., Aue, W. R., & Claridge, M. (2018). Release from output interference in recognition memory: A test of the attention hypothesis. *Quarterly Journal of Experimental Psychology, 71*(5), 1081–1089.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology, 53*(1), 1–13.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134–140.

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica, 142*(1), 1–5.

Forrin, N. D., MacLeod, C. M., & Ozubko, J. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40*(7), 1049–1055.

Gardiner, J. M., Craik, F. I. M., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior, 11,* 778–783.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). *Bayesian data analysis* (3rd ed). Boca Raton, FL: CRC Press.

Hacker, M. J. (1980). Speed and accuracy of recency judgements for evens in short-term memory. *Journal of Experimental Psychology: Human Learning and Memory, 6*(6), 651–675.

Hockley, W. E., & Cristi, C. (1996a). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition, 24*(2), 202–216.

Hockley, W. E., & Cristi, C. (1996b). Tests of the separate retrieval of item and associative information using a frequency-judgement task. *Memory & Cognition, 24*(6), 796–811.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33,* 36–67.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Kılıç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to percieve the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology, 92,* 65–86.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica, 59*(1), 57–69.

Koop, G., 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.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course.* Pittsburgh, PA: Carnegie Mellon University Press.

MacLeod, C. M., Nigel, G., Hourihan, K., Neary, K., & Ozubko, J. (2010). The production effect: Deliniation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 671–685.

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

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.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology, 86*(2), 263–267.

O'Neill, M. E., Sutcliffe, J. A., & Tulving, E. (1976). Retrieval cues and release from proactive inhibition. *American Journal of Psychology, 89*(3), 535–543.

Patterson, K. E., Meltzer, R. H., & Mandler, G. (1971). Inter-response times in categorized free recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 417–426.

Raaijmakers, J. G., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H Bower (Ed.), *The psychology of learning and motivation* (Vol. 14, pp. 207–261). New York, NY: Academic Press.

Raaijmakers, J. G., & Shiffrin, R. M. (1981a). Order effects in recall. *Attention and Performance, 9,* 403–415.

Raaijmakers, J. G., & Shiffrin, R. M. (1981b). Search of associative memory. *Psychological Review, 88*(2), 93–134.

Regenwetter, M., & Robinson, M. M. (2017). The construct-behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review, 124*(5), 533–550.

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

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

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.

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*(5), 511–524.

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

Silvapulle, M. J., & Sen, P. K. (2011). *Constrained statistical inference: Inequality, order, and shape restrictions*. Hoboken, NJ: Wiley Blackwell.

Smith, A. D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*, 400–408.

Tulving, E., & Arbuckle, T. Y. (1963). Sources of intertrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *1*, 321–334.

Tulving, E., & Arbuckle, T. Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology*, *72*(1), 145–150.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague. *Journal of Memory and Language*, *50*, 289–335.

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*, 440–445.

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88.