Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research report

# Does inhibition cause forgetting after selective retrieval? A reanalysis and failure to replicate

*Kevin W. Potter*[*], *Lucas D. Huszar and David E. Huber*

*University of Massachusetts Amherst, Amherst, MA, USA*

ARTICLE INFO

*Article history:*
Received 16 October 2017
Reviewed 23 Nov 2017
Revised 2 February 2018
Accepted 22 March 2018
Action editor Chris Chambers
Published online 4 April 2018

*Keywords:*
Memory
Retrieval induced forgetting
Replication

ABSTRACT

Retrieval practice can produce forgetting, but it remains unclear using only behavioral data whether this forgetting is caused by targeted inhibition versus interference. Therefore, Wimber et al. (2015) used pattern classifier analyses of fMRI data to track individual memories in a novel variant of retrieval induced forgetting. After initial learning, people recalled target images across selective retrieval practice trials, and cortical activity patterns gradually became more similar to those evoked by the target pictures (i.e., pattern enhancement) and less similar to those evoked by competing pictures (i.e., pattern suppression). The key question was whether this inhibition of competing memories would cause forgetting. Wimber et al. found a significant forgetting effect ($p < .01$) on a subsequent forced choice picture recognition test, with lower accuracy for competitors than for baseline items. Because fMRI data is correlative, a causal interpretation of the data would require, at a minimum, more forgetting following cortical pattern suppression (as occurred for competitors) than cortical pattern enhancement (as occurred for targets). The interaction necessary to reach this conclusion was significant ($p = .041$). However, reanalyzing the original data revealed that the interaction depended on the decision to code missing responses as equivalent to choosing the wrong picture. Even if missing trials reflected memory failures, at worst they would produce 50/50 guessing, rather than an error every time. Treating these trials as missing, or setting them to chance performance, resulted in no reliable forgetting difference between competitors and targets. Because this might reflect inadequate statistical power, we undertook two replication attempts of the behavioral paradigm, failing both times to observe more forgetting for competitors than targets. In fact, we failed to find any forgetting at all. We conclude that the study of Wimber et al. does not support the conclusion that forgetting is caused by targeted inhibition.

© 2018 Elsevier Ltd. All rights reserved.

[*] *Corresponding author*. University of Massachusetts Amherst, Amherst, 01003, MA, USA.
E-mail address: kevin.w.potter@gmail.com (K.W. Potter).

https://doi.org/10.1016/j.cortex.2018.03.026
0010-9452/© 2018 Elsevier Ltd. All rights reserved.

## 1. Introduction

Over a lifetime, many things are learned but then forgotten. For instance, after moving, you receive a new phone number, and the repeated retrieval of this new phone number seems to eliminate the memory of your old phone number. It is widely accepted that learning something new (e.g., your new phone number), can cause greater forgetting as compared to a baseline situation without new learning (e.g., for instance, if after moving you retained your old number, but did not use it for a period of time). Why does practice retrieving one thing at the expense of another (i.e., selective retrieval) cause forgetting? Do you actively suppress your old phone number to clear the way for your new phone number (i.e., targeted inhibition)? Or does the recently-practiced new phone number block access to the old phone number (i.e., interference)?

Since the 1970s, formal memory models have explained forgetting through interference (e.g., Hintzman, 1986; Murdock, 1982; Raaijmakers & Shiffrin, 1981). These models, and modern versions of them, explain many results from behavioral studies of memory. However, the last few decades have seen support for an intriguing theoretical alternative: in some circumstances, particularly as a result of retrieval practice, a competing memory might be actively suppressed, and the lingering consequences of this suppression cause forgetting for that memory (Anderson, Bjork, & Bjork, 1994). This forgetting via targeted inhibition is said to be selective (i.e., uniquely applied to the competing memory) because forgetting is observed regardless of the cues used to prompt retrieval (Anderson & Spellman, 1995). The cause of selective retrieval forgetting has been fiercely debated (e.g., Raaijmakers & Jakab, 2013), and interference models have been proposed to explain results previously explained by targeted inhibition (Jonker, Seli, & MacLeod, 2013; Tomlinson, Huber, Rieth, & Davelaar, 2015). Thus, there is something of a stalemate between the theoretical alternatives based on the behavioral data.

Neural data indicates that some form of unlearning occurs, such as with the long-term depression of synapses (Richards & Frankland, 2017). However, the weakening of synaptic connections is not necessarily the same thing as the targeted inhibition of memories; the function of synaptic weakening might be equivalent to learned interference in a distributed representation (Norman & O'Reilly, 2003). Nevertheless, neural data can be used to distinguish between the different functional accounts of forgetting. The study of Wimber, Alink, Charest, Kriegeskorte, and Anderson (2015) took this approach, tracking neural responses over the course of new learning to determine: 1) if competing memories are down-regulated in some manner in the course of retrieval practice; and 2) whether this down-regulation causes subsequent forgetting. We now consider this study in detail as our findings call into question some of the conclusions reached by the original authors.

The study of Wimber et al. (2015) used a novel variant of Retrieval Induced Forgetting (RIF: Anderson et al., 1994). Fig. 1 presents a brief visual summary of the experimental design. In this study, after pre-exposure to the entire set of pictures, participants learned word–picture pairs with a subset of the pictures. The pictures fell into three categories, showing either faces, scenes, or objects. The words were randomly chosen (unrelated to the pictures). This initial learning was assessed by showing a previously studied word and asking participants to name the corresponding picture. After naming the picture, visual memory of the picture was tested with a forced choice between the picture and a highly similar picture depicting the same object, face, or scene. For instance, as seen in the first row of Fig. 1, after learning to associate the word ‘sand’ with a picture of Marilyn Monroe, a participant would practice recalling Marilyn's name in response to a memory probe with ‘sand’ and then complete a practice two-alternative forced-choice (2AFC), attempting to choose the correct picture of Marilyn Monroe. These word–picture pairs were tested twice, with accuracy feedback following each test trial (a green box would highlight the correct picture). Next, participants completed an additional study/test session with the same words once again, but each word was now paired with a novel picture drawn from a different category of pictures. For example, in the second row of Fig. 1, participants might study the word ‘sand’ with a picture of a hat. Thus, this first stage of the experiment involved three study/test cycles, with the first two establishing strong learning for a first picture while the third established weaker learning for a competing picture from a different category.

This initial learning was followed by selective retrieval practice, inside the MRI scanner, in which participants were shown three fourths of the cue words (the remaining fourth was held out to provide baseline conditions). Each selective retrieval trial presented a cue word and participants were instructed to create a mental image of the first picture (e.g., Marilyn Monroe) rather than the second picture (e.g., a black hat). Thus, the first picture was the ‘target’ and the second picture was the ‘competitor’ that might inadvertently intrude during this task. Following an interval of 4 s in which to create this mental image, participants gave a key press indicating the category of the target picture (face, scene, object, or “unknown”). If a response was made within 1.5 s, they received feedback (this feedback indicated the correct category but did not indicate the specific picture). Indeed, during the first cycle of selective retrieval trials, participants often selected the competitor category by mistake. However, across four cycles of selective retrieval, accuracy improved as the propensity to recall the competitor category declined.

Selective retrieval practice was followed by a final recognition test with a forced choice between two pictures (the same forced choice between highly similar picture pairs as occurred during initial testing). The competitor pictures, and their corresponding baseline pictures, were tested in a block of trials before testing the target pictures, and their corresponding baseline pictures. The instructions for the final task asked subjects to respond as quickly as possible, and to guess if they could not remember. In the RIF literature, the behavioral test of interest for this final task is whether there is forgetting specifically for competitor images that were assigned to the selective retrieval condition, assessed by comparing the difference in performance between these images and the competitor images assigned to the baseline condition. Indeed, Wimber et al. found significant forgetting

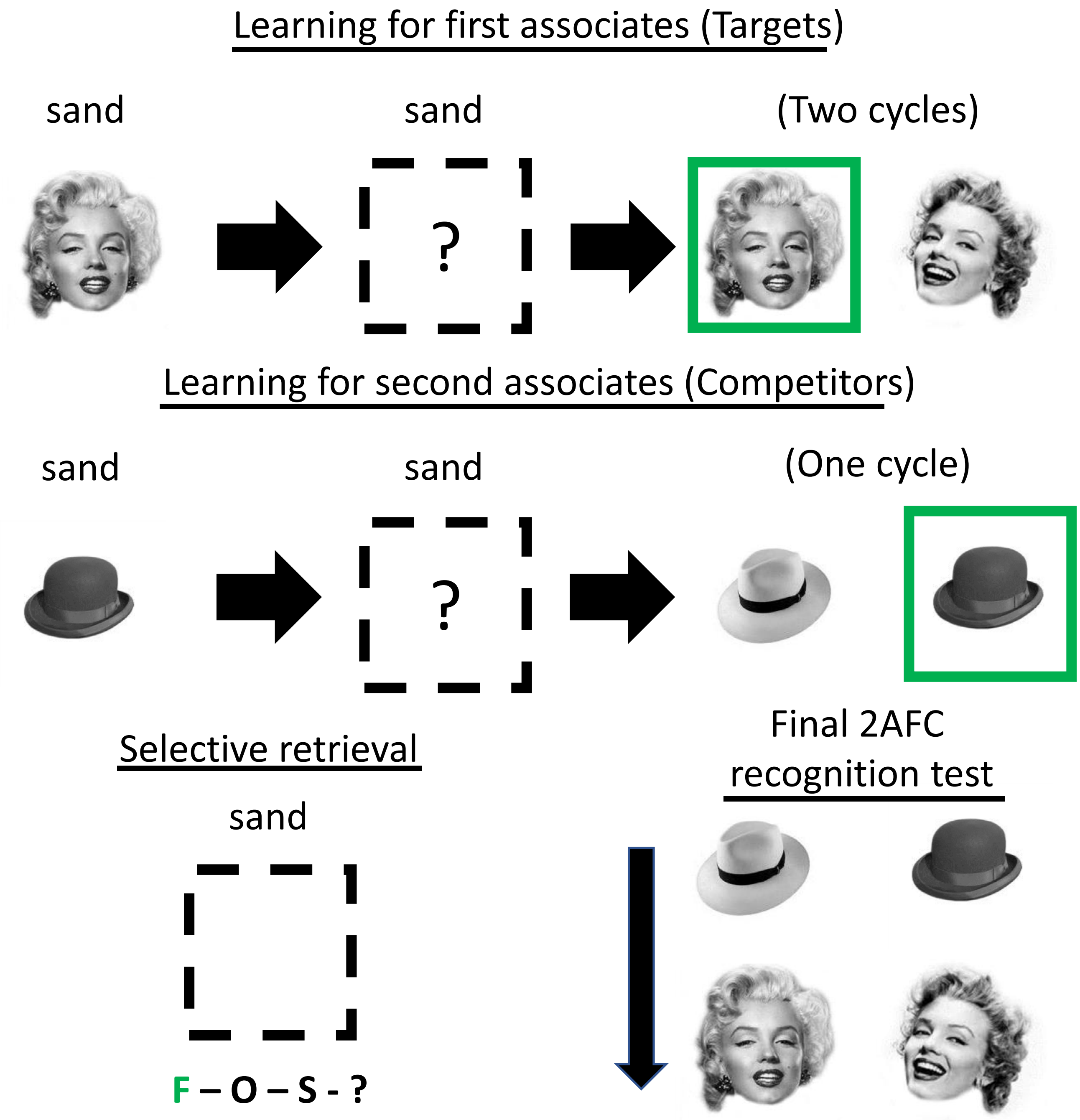
Learning for first associates (Targets)
sand
sand
(Two cycles)
?
Learning for second associates (Competitors)
sand
sand
(One cycle)
?
Selective retrieval
sand
F – O – S - ?
Final 2AFC recognition test

Fig. 1 – Visual summary of the experimental design used by (Wimber et al., 2015). Participants learned word–picture pairs and were instructed to create vivid visual memories (e.g., by having the images interact with the words). After learning, they completed an initial test phase. On each test trial, they engaged in verbal recall, speaking out loud the name of the picture associated with the presented cue word. Next, they attempted to pick the correct picture in a choice between two highly similar pictures showing the same face, scene, or object. Following each response, the correct picture was highlighted in green. Study and testing was broken up into blocks of 24 word–picture pairs. In the first block they completed two cycles for pictures assigned to be targets, and in the second block they completed a single cycle for pictures assigned to be competitors, with both blocks using the same cue words (e.g., the cue word might be associated with a face in block one and then an object in block two). Next, they completed selective retrieval practice (cued categorization) with 75% of the word–picture pairs while undergoing fMRI scanning. On each trial they saw the cue word and were given 4 s to visualize the picture that had been studied with that cue word in the first block. After visualization, they indicated the category (face, object, scene, or unknown) of the picture and then the correct category was highlighted in green. Finally, participants carried out force-choice recognition testing between the same picture pairs used during initial testing. No accuracy feedback was given during final tests and participants were given 3.5 s to respond. This final recognition was blocked, with testing of competitor pictures occurring before target pictures.

for competitors based on the accuracy data. Additionally, there was a 2 × 2 interaction, significant at $p = .041$, supporting a conclusion of more forgetting for competitors than targets. Somewhat surprisingly, the difference for targets also indicated forgetting, although this difference was not significant.

After this final recognition, participants viewed the same pictures once again, presented one at a time, while fMRI data were collected. This provided an item-by-item voxel pattern for different cortical regions, with these patterns used to assess the fMRI responses collected across the four trials of selective retrieval practice. In previous work, memory was assessed categorically (Kuhl, Rissman, Chun, & Wagner, 2011), but this approach was more fine-grained, asking whether specific memories were suppressed. Supporting this conclusion, across the four trials of selective retrieval, the voxel pattern elicited during the imagery task became progressively less similar to the voxel pattern evoked when viewing the competitor picture (this was true for visual areas and the hippocampus). At the same time, this voxel pattern became progressively more similar to the voxel pattern evoked when viewing the target picture. For visual areas, the suppression effect became significantly negative, indicating that the mental image created during retrieval was less like the competitor as compared to a different image drawn from the same category as the competitor.

It is tempting to consider the below baseline voxel pattern a direct measure of inhibition. However, this needn't be the case. For instance, a below baseline response could arise from avoiding thoughts of the competitor (while allowing thoughts of other pictures from the competitor category). The crucial question was whether this voxel pattern suppression caused subsequent forgetting. If so, this would support a targeted inhibition account of the observed neural activity. Thus, the key result hinged on the behavioral outcome (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). To establish that inhibition (as evidenced by pattern suppression) causes forgetting, one would want to know, at a minimum, that pattern suppression produced more forgetting than its opposite: pattern enhancement. The fMRI results revealed pattern suppression for competitors and pattern enhancement for targets, and so the interaction between item type (targets *vs* competitors) and selective retrieval condition (cued *vs* baseline) is necessary to support the conclusion that pattern suppression caused more forgetting than pattern enhancement. In the absence of such an interaction, one is left with the conclusion that pattern change caused forgetting (including pattern enhancement). Such a conclusion would not support forgetting through targeted inhibition; if the data cannot distinguish between the behavioral outcome following pattern suppression and the behavioral outcome following pattern enhancement, then the results do not support a causal role for pattern suppression.

Regarding the importance of the statistical interaction, it should be noted that a reliable interaction has not been deemed necessary in the RIF literature for establishing that forgetting occurred. However, the establishment that forgetting occurred is not the same as establishing the cause of forgetting. By analogy, consider a hypothetical study seeking to establish whether smoking causes lung cancer. Study participants are tracked over time providing initial baseline cancer rates and then cancer rates at the end of the study. With a simple comparison, suppose that the study finds that the rate of lung cancer increased for those who smoked during the period of the study. Furthermore, the same study failed to find a significant increase in lung cancer for non-smokers. At this point, it would be tempting to conclude that smoking causes lung cancer, but, upon further examination, it is revealed that the significant increase for smokers was supported by a statistical test with $p = .049$ (just under the .05 significance level) whereas the apparent lack of increase for non-smokers was based on $p = .051$. This problem is often referred to with the adage that the difference between significant and non-significant is not necessarily significant (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). If both groups reveal a numerical increase in cancer rates, this might just reflect the passage of time (i.e., cancer rates increase with age) that just happened to produce a slightly lower $p$-value for one group than the other. To conclude that smoking was a cause, it needs to be established that the increase for smokers was reliably greater than the increase for non-smokers. This difference of differences in a 2 × 2 design (pre-/post-crossed with smoker/non-smoker) is a statistical interaction. Returning to the Wimber et al. study, this analogy should make it clear that the conclusion regarding the cause of forgetting (i.e., that forgetting was caused by cortical pattern suppression) critically hinges on the $p = .041$ interaction.

In light of these considerations, we obtained the original behavioral dataset to consider, for instance, whether reaction times during recognition testing could shed light on these forgetting effects. In the course of our investigation, we found that some of the trials labeled as errors did not include a reaction time. We contacted the original authors, who confirmed that on some trials subjects failed to make a decision within the 3.5 s allowed for forced choice recognition, and that when this occurred, the trial was considered an error, equivalent to choosing the wrong picture.

In cued and free recall tasks it is standard practice to label the failure to respond as an error, equivalent to responding with the wrong answer. This is a sensible treatment of the data under the assumption that a failure to respond within the allotted time indicates that the subject cannot remember the answer, in which case the best they could possibly do would be to randomly guess. If the correct answer is a particular word, then the probability of randomly guessing that word from all known words is vanishingly small. In other words, the missing trial is replaced with the guessing probability of zero. However, the task used by Wimber et al. was not a recall task but rather two-alternative forced choice recognition. In this case, the guessing probability was 50% rather than 0%. Thus, the decision to replace these trials with 0% values did not respect the accuracy scale, which was lower bounded at 50% in the absence of any memory.

We are unaware of any memory process that causes people to reliably choose the wrong answer in a forced-choice task, as opposed to guessing randomly. Nevertheless, to assess this possibility, we performed an analysis of the reaction time data. If these missing trials were indicative of a choose-the-wrong-answer memory process, we would expect to see a reaction time distribution for error trials that extended out to the 3.5 s response deadline, with the missing trials corresponding to the chopped off tail of the error distribution. Instead, the data revealed something quite different; the very slowest observed error trial across all conditions and all subjects was nearly one whole second before the response deadline. Thus, these missing trials did not appear to be anything like the other error trials. Given that the missing trials did not appear to be part of the error distribution, they are best eliminated from the data analyses. We simply do not know how people would have responded on these trials if they had been given more time. It is even possible (and perhaps likely) that these missing trials would have produced better performance than the other trials if the reason that subjects missed the response deadline was because they were engaged in an in-depth analysis of the small visual differences between the two pictures.

The instructions for the recognition task stated “This task is about speed!” and “Select the correct picture as fast as possible, and guess if you cannot remember.” Despite these instructions, subjects failed to respond on 4.7% of the trials. This may seem like a small proportion but our analyses revealed that these missing responses played an important role in producing the reported results. We report evidence that this occurred because of the unbalanced design in which there were three times as many competitor trials (54 for each subject) as compared to the corresponding baseline trials (18 for each subject). Considering that forgetting is always assessed relative to baseline, the low trial count for the baseline condition makes the measure of forgetting highly variable. Our analyses revealed similar proportions of missing trials for competitors versus targets (5.2% *vs* 4.9%). However, the proportions of missing trials for the corresponding baseline conditions was quite different, with 2.7% missing for the competitor baseline versus 4.6% for the target baseline. Thus, because of the baseline trials, the decision to code these missing trials with zeros biased the results in favor of more forgetting for competitors than for targets. Although these numbers are small, the magnitude of the difference of differences for selective forgetting was also small (i.e. the extent to which there was more forgetting for competitors than targets).

We reanalyzed the original data, removing missing responses rather than labeling them as errors. We found that across all methods for assessing an interaction between target/competitor status and cued/baseline condition, the interaction was below acceptable levels of reliability. Furthermore, this was the case even if the missing trials were replaced with chance performance (50%) rather than labeling them as errors. If this interaction does not hold, it cannot be said that there was more forgetting for competitors than for targets, and without being able to reach this conclusion, it cannot be said that pattern suppression (but not pattern enhancement) caused forgetting.

It is important to note that a failure to find a reliable effect does not necessarily provide evidence against that effect. Instead, it may be that the study of Wimber et al. (2015) did not include sufficient statistical power to reach this conclusion, particularly in light of the missing trials, and in light of the unbalanced design. For this reason, we sought to replicate the behavioral results from this paradigm.

Our first replication attempt used a balanced design in a shorter version of this paradigm. We used the same pictures and words as the original study and we randomly assigned pictures to conditions. Unfortunately, this random assignment was the same for all subjects because the random number generator used to order stimuli in the experiment was set to the same seed. This did not introduce a confound (assignment was random rather than systematic) but made analyses difficult as this produced a between-items design, similar to a typical linguistics experiment. As reported in the supplementary materials, we used a mixed-effects analysis that properly addresses the role of item variability in a between-items design (Note the mixed effects models cannot parse out possible order effects introduced by a failure to counterbalance stimuli). Considering that the RIF effect is fairly well-established even when the final test is recognition (for a review, see Spitzer, 2014) we were surprised when this initial replication attempt failed to produce any forgetting effects, let alone more forgetting for competitors than targets. However, in light of this failure to counterbalance the stimuli, we next focus on our subsequent two attempts to replicate the reported forgetting effects, which used a different random assignment of stimuli for each subject.

Our second replication attempt used a design that was closer to the original study by collecting more data per subject, and by inducing a context shift between initial learning and retrieval practice (such as would occur when learning picture-word pairs outside the scanner followed by retrieval practice inside the scanner). This study randomly counterbalanced the assignment of items to conditions across subjects, but did not use a blocked design for the final recognition memory test. Instead, the order of competitor versus target images were intermixed. As with our initial replication attempt, this study failed to produce any forgetting, let alone more forgetting for competitors than targets.

We finally conducted a third highly powered pre-registered replication, this time matching the blocked design used by Wimber et al. for the final recognition test. Among our attempts, this replication was the closest to the original study. Furthermore, we had double the subjects of the original study, allowing us to identify sub-groups that better matched the performance of the original subjects. Nonetheless, with both our pre-registered and sub-group analyses, we again failed to find any forgetting, let alone more forgetting for competitors than targets.

Here we report on our second and third replication attempts and our analyses of the original data. Our goal is to correct the publication record, demonstrating: 1) that this novel variant of the retrieval induced forgetting paradigm does not appear to produce more forgetting for competitors than targets, and in fact does not appear to produce any forgetting at all; and 2) the original dataset does not support the conclusion that cortical pattern suppression caused more forgetting than pattern enhancement.

## 2. Method

In Fig. 1 we introduced the design used by Wimber et al. (2015), and we refer the interested reader to that study for additional procedures details. Below, we compare and contrast the original design against our replication attempts. Any unmentioned procedural details (e.g., the initial picture familiarization stage) were kept the same as the original study to the best of our knowledge, based on the reported procedures of the original study and our reanalyses of the original data. Our preregistration script may be found at https://AsPredicted.org/u5z76.pdf. The experimental materials, scripts, and analyses for all three replications are freely available at https://github.com/rettopnivek/Wimber_et_al_replication_3.

There are four important procedural details that were not reported in the original Wimber et al. (2015). For completeness, we present these details here. First, as mentioned in the introduction, Wimber et al. labeled missing trials from the final forced choice recognition test as errors, equivalent to choosing the wrong picture; during each final forced choice recognition trial, subjects were given 3.5 s to respond, and if no response was given within this time period, the experimental program automatically moved on to the next test trial. Second, the final recognition test list was blocked, testing all of the competitor picture pairs (and the corresponding baseline picture pairs) before testing the target picture pairs. This has become a standard approach in the RIF literature, in the attempt to avoid alternative explanations of forgetting for competitors based on output interference from already retrieved targets. Third, although the supplementary methods indicated that the assignment of picture pairs to conditions 'was counterbalanced such that across participants, each picture equally often served as a target, competitor and baseline item.', our analyses found that this was not the case, raising the possibility of unacknowledged item effects. To counterbalance the stimuli, Wimber et al. assigned picture pairs equally often to be first associates (i.e., target pictures which were initially studied and tested twice), versus second associates (i.e., competitor pictures which were initially studied and tested just once). However, the original study had an inherently unbalanced design, with 54 of the word cues appearing during selective retrieval, while only 18 word cues were held out from selective retrieval to provide each of the corresponding baseline conditions. In other words, there were three times as many target/competitor recognition tests as compared to the corresponding baseline recognition tests. Because of this, it was impossible to counterbalance assignment of pictures to the baseline and selective retrieval conditions. Out of the 24 subjects, each picture was assigned to the selective retrieval condition for 18 of the subjects on average, but that same picture was assigned to the baseline condition for only 6 of the subjects on average. Furthermore, these are just average numbers and there was a great deal of variability across pictures. For instance, for the baseline conditions, the number of times each picture was actually assigned to be a competitor baseline item or a target baseline item ranged from 2 to 10. Finally, it was not reported that the final recognition instructions emphasized speed, asking subjects to respond as quickly as possible and to guess when they did not know the correct answer.

### 2.1. *Participants*

#### 2.1.1. *Original design*

Wimber et al. recruited 24 subjects from a volunteer panel at the MRC Cognition and Brain Sciences Unit. The sample consisted of 20 females and 4 males, ranging from 20 to 32 years of age with a mean of 24.2.

#### 2.1.2. *Second replications*

For the second replication attempt, 35 subjects were recruited via word of mouth and leaflets from the university campus. Subjects received $20 for their participation. The replication was approved by the UMass Amherst IRB and written informed consent was obtained from all subjects. We originally collected data from 25 subjects. However, 10 subjects received instructions that flipped the key assignments for two responses during the selective retrieval stage. Therefore, we ran an additional 11 subjects. In the main text, we report the results for the 25 subjects that received the correct instructions. In our supplementary material, we report the results for all 35 subjects, along with an evaluation of the impact of the incorrect instructions.

For our second replication, we removed certain subjects and responses based on a priori criteria. First, we removed subjects whose overall accuracy during the practice test was at or below chance performance. To determine this, we fit each subject's accuracy data with a simple beta-binomial model with uniform priors, and we excluded subjects whose 95% credible intervals included (or fell below) a probability of .5. Also, we removed subjects who made an excessive number of time-out responses (i.e., trials where they failed to make a response within a set time limit of 4 s for the selective retrieval phase, and 3.5 s for the recognition memory phase). If 25% or more of the total number of responses were time-out responses, that particular subject was removed (although this cut-off is somewhat arbitrary, we note that it was made in advance of data collection). Finally, we removed any time-out responses (whereas Wimber et al. labeled time-out responses as incorrect). After excluding the 10 subjects with reversed instructions, we had 22 subjects who met the inclusions criteria for the second replication. We note that if these exclusion criteria are applied to the data from Wimber et al., two subjects would have been excluded due to at-chance performance during the practice tests, leaving only 22 subjects for the original study as well.

#### 2.1.3. *Third pre-registered replication*

We recruited 53 subjects via word of mouth and leaflets from the University of Massachusetts Amherst campus, paying subjects $20 for their participation. The third replication was approved by the university IRB and written informed consent was obtained from all subjects. 4 subjects did not complete the study, and one subject provided no responses during the selective retrieval stage. This left us with 48 subjects with complete data, our desired sample size (as noted in the preregistration script). This sample size was chosen based on an a priori power analysis: we simulated data 1000 times from a hierarchical logistic regression model with a simple effect of forgetting for competitors. The original study reports an effect of 7%, and if missing data is trimmed, the effect is still 5%. However, we determined that with a sample size of 48 subjects, our power to detect an even smaller effect of forgetting for competitors of 4% was above 95%. Our sample consisted of 22 females and 26 males. The mean age was for our sample was 24.65 (SD = 5.21); the youngest subject was 18, while the oldest was 39. Note that for our third replication, we did not exclude subjects or trim any data based on the criteria from our previous replications. This approach is consistent with Wimber et al.'s original methodology.

### 2.2. Materials

Wimber et al. used 72 English words taken from the MRC linguistic database (http://www.psych.rl.ac.uk/). They selected words that had relatively low image ability and concreteness ratings, to avoid having the words themselves elicit mental images while subjects were in the scanner. Wimber et al. also used 144 picture pairs, with three sets of 48 picture pairs of well-known faces, scenes, and everyday objects. For example, subjects could see pictures of Marilyn Monroe's face at two different angles, two different soccer balls, or two different angles of the Acropolis. Pictures in the pairs were assigned to either be studied pictures or lure pictures, and this assignment was the same across all subjects (i.e., picture pairs are the base unit for item effects considering that the same picture was the lure picture for all subjects). Wimber et al. selected these pictures from a variety of databases and the internet, converted them to grayscale, rescaled them to cover the same visual angle, and stripped the background from the faces and objects pictures (but not the scenes). Wimber et al. kindly provided us with their stimuli, so for all three replication attempts we were able to use the same materials.

### 2.3. Design and procedure

#### 2.3.1. Second replication

We now discuss the procedure and design of our second replication attempt, focusing for brevity on how it differed from the original study. The second replication used the full stimulus set, but in contrast used a balanced design. 36 picture pairs were randomly assigned to each condition. This contrasted with the unbalanced design of Wimber et al., where 54 picture pairs were assigned to the selective-retrieval conditions and only 18 picture pairs were assigned to the baseline conditions. Also, Wimber et al. used a blocked final recognition test in which subjects were first tested on all of the competitor pictures before being tested on the target pictures. For the second replication, the order of target and competitor pictures was randomized during the final recognition test (because the nature of the blocked design was unreported, we assumed a mixed final test list, only learning of this blocking after our first two replication attempts). We also attempted to impose a context shift by having subjects complete the selective retrieval and recognition memory phases in an EEG testing room whereas the initial training phase occurred in a simple testing room in a different part of the building.

#### 2.3.2. Third pre-registered replication

We used the same stimuli and the blocked, unbalanced design used in the original study. However, due to our large sample size, we randomly assigned picture pairs to each of the four conditions instead of counterbalancing assignment to first associates versus second associates. Unlike Wimber et al., we statistically controlled for potential item effects. As per our previous two replications, we used a purely behavioral design. Subjects did not undergo fMRI scanning during the selective-retrieval phase. It should be noted that a subject's performance can consistently differ when undergoing fMRI scanning. For example, van Maanen, Forstmann, Keuken, Wagenmakers, and Heathcote (2016) found that subjects exhibited slower motor responses and attenuated attentional focus in the scanner (the balance between motor slowdown and poorer focus was task specific). Unfortunately, if the original forgetting effect can only be replicated when subjects are in a scanner, then it can hardly reflect the mechanisms underling everyday memory failures.

Unlike our second replication, subjects completed the entire task (which took approximately 2 h) in a single room. Also, we had subjects complete a 5-point Likert scale at the end of the study indicating to what extent they visualized the images they were supposed to recall during the selective-retrieval stage. The response categories were: 1) I only thought of the name of the category, 2) I very rarely mentally visualized the image, 3) sometimes I was able to create a mental image but not always, 4) I typically was able to create mental images, and 5) on nearly every trial I was able to create detailed mental images.

## 3. Analyses

In this section, we describe our analytic approach. We used a different statistical framework and set of tools to analyze our data compared to Wimber et al., so we first describe their approach and our motivation for using different analyses. We then focus in detail on our analyses for assessing the results of the selective retrieval and final recognition memory stages of our replication attempts. We also further report on our a priori power analysis for our pre-registered design, and note the priors we used in our Bayesian analyses. However, for ease of comparison, we also report Wimber et al.'s original analyses as applied to our data sets. Critically, it did not matter how we analyzed our results and in every case, our replication attempts failed to produce any forgetting, let alone more forgetting for competitors than for targets.

### 3.1. Original analyses

The primary behavioral results of interest in Wimber et al.'s experiment involved the selective retrieval stage, and most critically, the final recognition memory test. To analyze these data, Wimber et al. used the standard suite of statistical tools known to psychologists.

In the selective retrieval task, subjects could make one of four possible responses for the subset of 54 word cues that they saw. Subjects could correctly pick the category of the target picture (a hit), incorrectly select the category of the competitor picture (an intrusion), pick the third, unrelated category (an error), or indicate that they did not know the answer (unknown). Furthermore, subjects saw each cue word 4 times, allowing for an analysis of the change based on repeat presentations.

Wimber et al. used a paired samples *t*-test to determine that subjects had significantly more intrusions than errors ($t_{23} = 6.53$, $p < .001$). Additionally, they verified via a one-way repeated measures ANOVA that the proportion of intrusions varied significantly over repetitions ($F_{3,69} = 21.8$, $p < .001$), exhibiting a linear decline ($F_{1,23} = 55.4$, $p < .001$).

In the final recognition memory task, subjects had to select the studied picture instead of a similar looking lure picture. There were four conditions in this task: pictures could either be targets or competitors (i.e., first versus second associates), and cue word associated with each picture could have been shown during the cued categorization task (selective retrieval) or not (baseline). Wimber et al. examined the proportion of correct responses over these four conditions.

Wimber et al. conducted a 2 × 2 repeated measures ANOVA on the proportion correct (after rounding these values to two decimal places) with two factors: associate type (target *vs* competitor) and assignment type (selective retrieval *vs* baseline). The authors found a significant interaction ($F_{1,23} = 4.70$, $p = .041$). Though it was not reported, there was also a main effect of assignment type ($F_{1,23} = 13.61$, $p < .001$), indicating a general forgetting effect when collapsing across targets and competitors. The authors then conducted planned comparison paired samples *t*-tests comparing the difference between selective retrieval and baseline conditions for performance with competitors and targets respectively. They found that performance for competitor pictures that underwent selective retrieval had significantly lower performance compared to performance for competitors in the baseline condition ($t_{23} = 4.91, p < .001$), with an average proportion correct of .752 versus .821, respectively. In contrast, performance did not differ significantly between selective retrieval and baseline conditions for target pictures ($t_{23} = 0.57$, $p = .713$), with an average proportion correct of .786 versus .797, respectively.

### 3.2. *Motivation for a different approach*

Our choice of statistical models differs from those used in the original analyses by Wimber et al. for their behavioral data. More specifically, because there was an a priori expectation that we would find a forgetting effect, we employed a Bayesian framework such that the results of Wimber et al. served as priors for our study, biasing our statistical analyses in favor of the conclusions reached by Wimber et al. In addition, this allowed us greater statistical power by appropriately addressing the bounded, count nature of the data.

Wimber et al. primarily focus on the proportion of intrusions and correct choices when analyzing their behavioral data. However, the t-test and ANOVA both assume the data being analyzed are unbounded and continuous. In contrast, proportions are bounded between 0 and 1 and based on finite count data, meaning that they are not continuous. Therefore, as noted by Jaeger (2007), the probability model underlying the *t*-test and ANOVA can incorrectly assign probability mass to impossible values when dealing with proportions (i.e., values that fall below 0 or above 1). This issue is especially problematic when subjects have proportions close to the extremes, a common occurrence in the current experimental paradigm. Furthermore, Jaeger notes that the variability of count data changes based on the underlying probabilities, with greater variability in the data occurring when probabilities are near .5. In other words, count data exhibits heteroscedasticity, violating further assumptions of the ANOVA approach.

Wimber et al. also used a limited set of test items, examining the same 144 picture pairs across all 24 subjects. Our analyses demonstrated substantial item effects, with some picture pairs producing reliably better recognition performance than others, regardless of experimental condition. For instance, it is easy to distinguish a scene of the Acropolis from its lure (viewed at a very different angle) relative to two very similar looking soccer balls. This fact, combined with the inherently unbalanced design used by Wimber et al., emphasizes a need to simultaneously control for both item and subject effects. However, a repeated measures ANOVA can only control for one effect or the other, but not both. Based on these concerns, we elected to use mixed effects logistic regression when analyzing the data for the final recognition memory test from our replications.

### 3.3. *Our analytic approach*

We relied on two types of statistical models when analyzing our replication attempts. First, we used a hierarchical categorical logit model (e.g., Stan Development Team, 2017c). This model is useful for categorical data with multiple response types, such as the cued categorization task with its 4 response categories (hits, intrusions, errors, and "unknown" responses). The model assumes that the data follow a categorical distribution (i.e., separate probabilities for each response type that sum to one), and uses a softmax link function (also known as the normalized exponential function) to estimate an associated set of unbounded, continuous parameters. We assumed separate parameters for every subject that were informed by a single set of group–level parameters. This type of model captures the bounded nature of the count data, the variability due to individual differences, and the inherent dependencies between response types (e.g., a high proportion of hits means a much lower proportion of intrusions, errors, and "unknown" responses).

We also used hierarchical logistic regression with mixed effects to analyze dichotomous count data. This model assumes the data follow a binomial distribution, and uses the log of the odds as the link function. The resulting unbounded, continuous parameter can be decomposed into the standard weighted sum of a set of predictors. We used this model to analyze linear trends in the cued categorization task, collapsing the 4 response types into a binary response (e.g., intrusions versus all other response types). We assumed a random effect for subjects, and used a planned contrast to test for a linear trend. In the cued categorization task, we were most interested in the trend on hits and intrusions across the 4 selective retrieval trials that tested the same cue. Logistic regression is very useful in this scenario because the intrusions in particular were close to the boundary of 0. We also used the hierarchical logistic regression to examine accuracy performance in the final recognition memory task. Here we assumed random effects for both subjects and for items (i.e., picture pairs).

We estimated parameters within the Bayesian framework. Compared to the frequentist approach, the Bayesian framework has a more coherent approach for testing hypotheses, and allows for a more intuitive interpretation of uncertainty in parameter estimates (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). Furthermore, the Bayesian framework is well suited for the estimation of hierarchical models. Finally, the Bayesian approach is inherently well-suited to replication attempts, as the posteriors for model parameters from the original analysis can be used as priors for the subsequent analyses of the replication. We conducted all analyses using the statistical software R (version 3.4.1, R Core Team, 2017). We estimated the hierarchical logistic regression models using the R package ‘rstanarm’ (version 2.15.3 Stan Development Team, 2017a), and we estimated the categorical logit models using custom scripts written with the R package ‘rstan’ (version 2.16.2 Stan Development Team, 2017b). Again, our analysis scripts are freely available at https://github.com/rettopnivek/Wimber_et_al_replication_3.

### 3.4. *Analyses for pre-registered design*

We have outlined the core statistical models we used to analyze our replication attempts. We now describe the specific details for the pre-registered analyses used with our third, final replication effort. First, when analyzing the final recognition memory task, we specified a confirmatory statistical model instead of a purely descriptive one. For the fixed effects in the hierarchical model, we specified an intercept term to represent performance in the baseline conditions for both targets and competitors as well as the selective-retrieval condition for targets. We specified a separate coefficient capturing the change owing to selective retrieval practice.

Hence, the statistical model specifically predicted the type of cross–over interaction needed to support Wimber et al.’s conclusion that cortical suppression, not enhancement, leads to forgetting. The model could not account for interactions or main effects of a different nature, providing a strongly confirmatory test of Wimber et al.’s original findings. Moreover, while the confirmatory model was designed to test for the interaction originally found in Wimber et al.’s data, rather than the simple comparison typically of interest in the RIF literature, it is important to note the confirmatory model can still provide an estimate of RIF that closely matches that of the simple comparison.

As noted earlier, we checked whether this confirmatory test had sufficient power to detect a forgetting effect. Fig. 2 reports the power our confirmatory model had to detect a forgetting effect of 4% or lower based on a sample size of 48 (solid line) or 24 (dashed line) subjects, determined from Monte Carlo simulations using 1000 repetitions (therefore estimates of power can vary by ± 0.02). As the figure shows, with our pre-registered sample size of 48 subjects, we had over 95% power to detect an effect of 4%, and we still had above 80% power to detect an effect of 3%. Recall that Wimber et al. originally reported an effect of 7% (5% after correcting for missing data). Therefore, we had a high-powered pre-registered replication design.

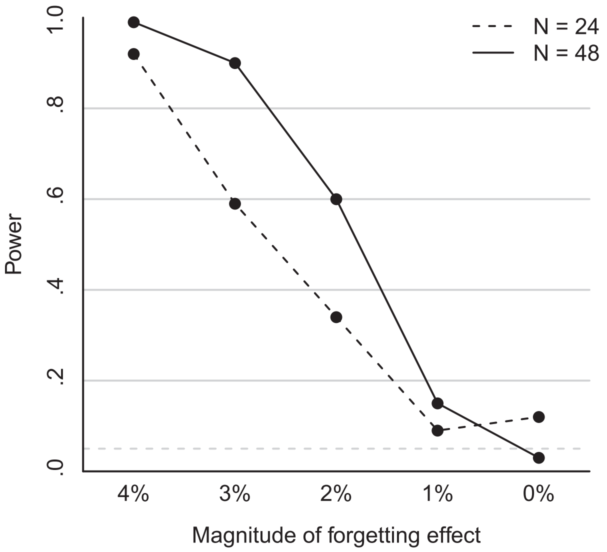

**Fig. 2 – The power to detect a forgetting effect of 4% or lower using a confirmatory mixed effects logistic regression model, computed via Monte Carlo simulations over 1000 repetitions. The solid line indicates the power for a sample size of 48 (the number of subjects for our pre-registered design) while the dashed line indicates the power for a sample size of 24 (the number of subjects from the original study). The degree of Monte Carlo error is about 2%.**

Because we estimated the statistical models within a Bayesian framework, it is important to note the priors that we used in our analyses. To specify the priors, we first reanalyzed Wimber et al.'s original data. Once we obtained the marginal posterior estimates, we set these as the new priors for the analysis of our replication. This biases the results to favor the findings of the original study, but it greatly increases our power to detect greater forgetting for competitors as compared to targets. We assumed normally distributed priors for the logistic coefficients for the fixed effects in our models, and we report the values using the format N ($\mu$, $\sigma$), where $\mu$ refers the mean and $\sigma$ the standard deviation. For the initial reanalysis, we placed a N (1.775, .3) prior on the intercept, and a N (-.3, .3) prior on the coefficient. The subsequent marginal posteriors (and new priors for the analysis of the replication) were N (1.57, .12) and N (-.28, .08). Note that these values differ from those reported in the preregistration script (which were based on a statistical model with an additional covariate for the effect of training). A common concern often raised regarding priors is that they allow for additional researcher degrees of freedom, which can inflate Type I error rates (Simmons, Nelson, & Simonsohn, 2011). Fortunately, because we used the posteriors from the original data and pre-registered the analysis of our third replication, this concern is not applicable for our situation.

There are multiple ways to define a successful replication. For example Nature Neuroscience, the journal in which

Wimber et al. results were published, defines a successful replication to be when the effects for the replication fall within the credible intervals of the original data (A. Arguello, personal communication, May 10, 2016). We therefore used two criteria to assess whether we successfully replicated Wimber et al.'s finding of forgetting for competitor images. First, we used a posterior predictive check, generating the range of plausible values for the average proportion correct in each condition using the posteriors from the reanalysis of the original data. If our replication data falls within the resulting credible intervals, this indicates that the original estimates can successfully predict new data from the same experimental design. Note that because the original study had a smaller sample size, these intervals will be wider and less precise. As noted in our preregistration script, another way to assess the success of our replication is to determine whether the marginal posterior for the coefficient describing the forgetting effect contains the original point estimate. However, this assessment is only appropriate if the statistical model is able to fit the replication data in the first place. Because we fit a confirmatory model, an important sub-step is to generate a posterior predictive check with the replication estimates and assess whether they actually fit the replication results.

## 4. Results

We first present our reanalysis of the data from Wimber et al., focusing on the appropriateness of labeling missing data as incorrect, and the reliability of the interaction between the type of associate (targets *vs* competitors) and condition (whether images underwent selective retrieval or not). Next, we report our findings for our first two replication attempts. We then present our results for our pre-registered replication, as well as the additional sub-group analyses we conducted.

### 4.1. *Reanalysis of original data*

One way to assess the appropriateness of coding missing data as errors is to examine how close error response times occurred relative to the deadline. If error responses were especially slow, then it is reasonable to expect that a proportion of these errors could extend past the deadline and be registered as missing responses. This means that there would be a gradually diminishing set of error response times that would abruptly be truncated at the 3.5 s response deadline. We can test this by taking the log-transform of the error responses times (to adjust for the fact that response time distributions are positively skewed), and then examine the distance in standard deviation units of the deadline from the mean of error response times for a particular subject. This can then be converted into the predicted proportion of responses to exceed the deadline via the standard normal cumulative distribution. This approach can over-estimate the distance from the deadline, but this over-estimation will be minuscule given that the slowest error response made by the subjects from the original study was 2.569 s. In other words, there was a gap of 931 ms with no trailing error responses before the deadline.

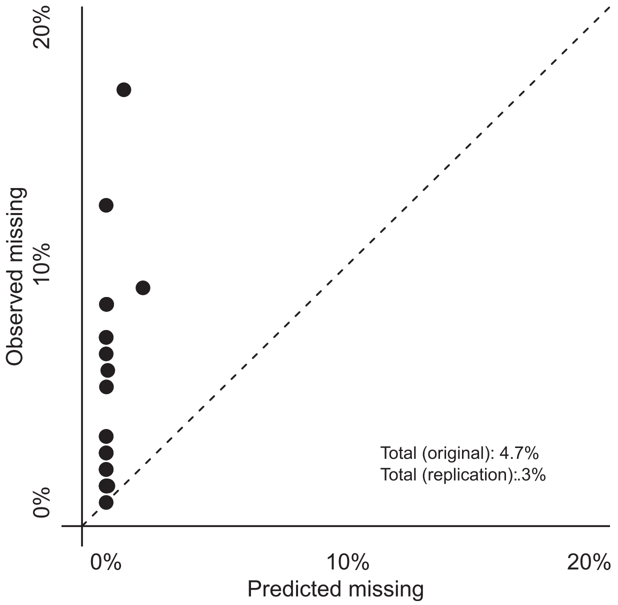

**Fig. 3 – The predicted versus observed percentages of missing data from Wimber et al.'s data for the final recognition memory task. The *x*-axis shows the predicted percentage of missing data for each subject, estimated from how far away the log of the deadline (3.5 s) was from their average log response time. The *y*-axis shows the observed percentage of missing data for each subject. Values above the dashed line indicate more missing data than predicted.**

Fig. 3 shows a scatter plot of the predicted percentage plotted against the observed percentage of missing responses for each subject. Points that fall above the dashed line indicate more missing responses observed relative to the predicted amount, under the assumption that missing trials would have been errors. As can be seen, the proportion of observed missing responses far exceeds the predicted amount based on error response times. Therefore, it more likely that the missing data reflect ancillary processes rather than a retrieval process that was destined to result in choosing the wrong picture. In contrast, our replication data had far fewer time-out responses (less than half of a percentage point).

#### 4.1.1. *Robustness of the interaction*

Having confirmed that the labeling of time-out trials as incorrect responses was inappropriate, our next question was how reliable the interaction term was if these time-out trials were instead treated as missing data. We examined the robustness of the interaction term using four different statistical tests, ordered in descending fashion to reflect the level of appropriateness for the current data. First, we report the results of the standard repeated measures ANOVA as applied to percent correct. Next, we report the results of a hierarchical logistic regression that controlled for subject effects. We then report the results for a hierarchical logistic regression that also incorporates item effects. Finally, we report the results of a hierarchical multinomial process model with subject and item effects.

The multinomial process model posits that subjects either recognize the correct picture, or they guess with 50% accuracy. This model therefore better accounts for the fact that with a forced choice test, the hypothetical 'floor' is 50% accuracy rather than 0%. The standard logistic regression approach assumes that the data follow a binomial distribution governed by the probability P(Correct), and the log-odds of this probability can be linearly decomposed. In contrast, with the multinomial process model, the probability governing the binomial distribution is instead

$$\mathrm{P}(\mathrm{Recognized}) + \frac{1}{2}[1 - \mathrm{P}(\mathrm{Recognized})] \quad (1)$$

It is then the log-odds of the probability P(Recognized) that is linearly decomposed.

We report the results of the four tests over the two datasets in Table 1. The first column lists the type of statistical model. The second column reports the type of scale (in percent units) that the model assumes for the data. This column makes it clear that the ANOVA and logistic regression approaches do not correctly bound the response scale for the forced choice test. Note that the logistic regression and multinomial process models have improved power to detect a reliable interaction, especially after including item effects, since they parse out additional variability in the data for which the original ANOVA did not control.

We report the significance of the interaction between associate type and condition for the original data in which missing data was labeled to be errors, versus the original data in which missing data is excluded from the analyses. Statistically significant results are marked with asterisks. When missing data were labeled as errors (rather than trimmed), one interesting finding is that when one switches from a repeated measures ANOVA to logistic regression, which assumes more appropriate boundaries for the percentage correct, the interaction is only marginally significant. This is because the relation between the underlying linear structure and percentage correct is non-linear, and only certain types of interactions are robust to these non-linear transforms. However, incorporating item effects and controlling for guessing via the multinomial process model recovers the significance of the interaction for the original data set. This again emphasizes how the multinomial process model in particular has improved power to detect a reliable interaction. Critically, though, if one trims out the missing data, the interaction is not significant regardless of the statistical test.

**Table 1 – The *p*-values for the interaction of associate and retrieval type.**

| Test | Scale (Percent) | Missing (error) | Missing (trimmed) |
|---|---|---|---|
| ANOVA | $-\infty$ to $+\infty$ | .041* | .106 |
| Logistic | 0 to 100 | .063 | .145 |
| Logistic (Items) | 0 to 100 | .053 | .129 |
| Multinomial process | 50 to 100 | .038* | .071 |

* Significant at the $\alpha = .05$ level.

### 4.2. *Second replication attempt*

Fig. 4 presents the results for our second replication attempt. Panel A shows the average response proportions (hits, intrusions, errors, and unknowns) and their associated uncertainty (based on the categorical logit model) for the cued category task during selective retrieval practice. In terms of learning during selective retrieval practice, we replicated the qualitative patterns from Wimber et al.'s study. On average, subjects correctly identified the category for the target image 71% of the time. Subjects also had more intrusions than errors (15% *vs* 9%, with a posterior *p*-value less than .001 for the difference). Note, however, that quantitatively speaking our subjects from the second replication made fewer hits on average compared to those from the original study, who on average identified the target 76% of the time. Furthermore, our subjects had a greater number of intrusions and errors relative to the original subjects, who on average only chose the competitor image 9% of the time, and made an error 2% of the time. Finally, our subjects on average picked "unknown" less of the time (7% compared to 13% for the subjects of the original study).

Panel B presents the trend analysis for hits and intrusions using the mixed effects logistic regression model. The plots on the left show the change in the proportion of hits and intrusions across the 4 selective retrieval trials that tested the same cue word, and the plots on the right show the slope of the linear trend estimated from the mixed effects logistic model. Error bars represent 95% credible intervals. Demonstrating that subjects learned during selective retrieval, there was a significant improvement in hits and a decline in intrusions across the 4 trials. Again, while subjects had fewer hits on average and more intrusions, we replicated the linear trend reported by Wimber et al., which they argued was consistent with the possibility that inhibitory control rendered competitors less interfering over cue repetitions.

Panels C presents the results for the four conditions of the final recognition memory task, averaged over subjects. The solid lines and filled points represent the average proportion correct for targets, whereas the dashed lines and empty squares represent the average proportion correct for competitors. Error bars represent the 95% credible intervals generated via simulations based on the posterior samples, indicating the range of plausible group means conditioned on the observed data. We found no interaction ($\beta = -.113$, posterior *p*-value $= .273$) nor forgetting for competitors ($\beta_B - \beta_{SR} = 0.004$, posterior *p*-value $= .511$).

Applying Wimber et al.'s original analyses to the data for the second replication leads to similar conclusions. The difference between the average proportion of intrusions and errors is statistically significant according to a paired samples t-test, $t_{21} = 5.461$, $p < .001$. The one-way repeated measures ANOVA applied to the proportion of intrusions over cue repetitions was significant, $F_{3,63} = 22.5, p < .001$, and also exhibited a linear decline, $F_{1,21} = 42.3$, $p < .001$. However, for the final memory task, the repeated measures ANOVA found no significant effects. For the main effect of associate, $F_{1,21} = .168$, $p = .686$, while for the main effect of condition, $F_{1,21} = .493$, $p = .49$, and finally, for the interaction, $F_{1,21} = 0.126$, $p = .726$.

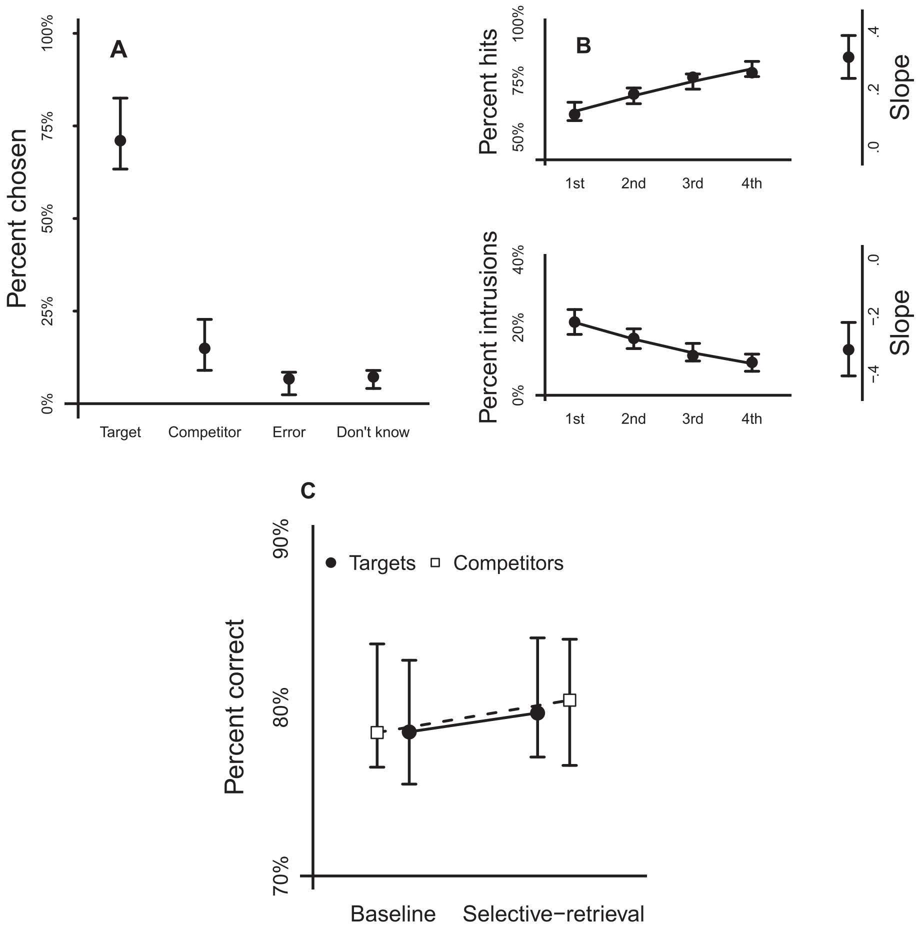

**Fig. 4 – Performance for the selective-retrieval phase and final recognition memory task for our second replication attempt. (A) The average percentage in which the categories for a particular picture type (target, competitor, incorrect, or unknown) were chosen collapsed over all selective retrieval trials. Error bars represent 95% credible intervals from a categorical logit model. (B) A trend analysis of the percentage of hits (correctly picking the target) and intrusions (incorrectly picking the competitor) over the 4 selective retrieval trials that presented the same cue word. 95% credible intervals for the slope of the linear trend from a logistic regression analysis are shown to the right. (C) Performance in the final recognition memory task averaged over the 22 subjects. The conditions for target images are denoted by filled circles and solid lines, while the conditions for competitor images are denoted by empty squares and dashed lines.**

We then ran the planned comparison paired sample *t*-tests. The comparison for competitors versus baseline was not significant, $t_{21} = -0.176$, $p = .862$, nor was the comparison for targets versus baseline, $t_{21} = -0.838$, $p = .411$.

We can also examine the significance of the interaction over progressively more powerful statistical tests and differing treatment of missing data, as we did with the data from the original study. As seen in Table 2, the interaction is not significant regardless of the test. Furthermore, coding missing responses as errors had no impact, which is unsurprising as there were only two time-out responses for the 22 subjects.

However, the experimental designs of our second replication attempt differed from the original study by using a randomly ordered rather than blocked final test list and by using a balanced design. In typical RIF designs, the blocked design is important to negate an alternative interpretation of the results in terms of output interference. We therefore focus on the results of our third, pre-registered design, which closely matched Wimber et al.'s original study.

**Table 2 – The *p*-values for the interaction of associate and retrieval type (Second replication).**

| Test | Scale (Percent) | Missing (error) | Missing (trimmed) |
|---|---|---|---|
| ANOVA | $-\infty$ to $+\infty$ | .726 | .726 |
| Logistic | 0 to 100 | .681 | .681 |
| Logistic (Items) | 0 to 100 | .631 | .631 |
| Multinomial process | 50 to 100 | .438 | .435 |

### 4.3. *Pre-registered replication attempt*

Fig. 5 reports the results of our pre-registered replication attempt (in black) compared against Wimber et al.'s original findings (in red). Panel A shows the average response proportions (hits, intrusions, errors, and unknowns) and their

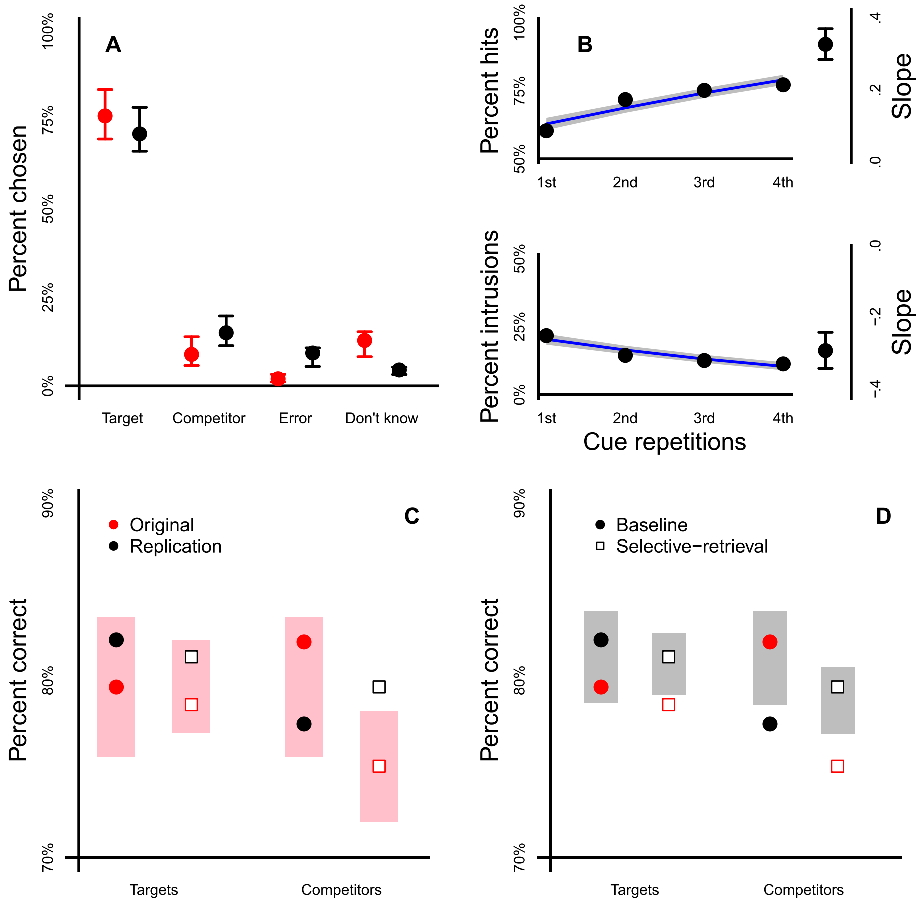

**Fig. 5 – Performance for the selective-retrieval phase and final recognition memory task for the pre-registered direct replication (in black). For easy comparison, the results from the original study are included (in red). (A) The average percentage in which the categories for a particular picture type (target, competitor, incorrect, or unknown) were chosen collapsed over all selective retrieval trials. Error bars represent 95% credible intervals from a categorical logit model. (B) A trend analysis of the percentage of hits (correctly picking the target) and intrusions (incorrectly picking the competitor) over the 4 selective retrieval trials that tested the same cue word. 95% credible intervals for the slope of the linear trend from a logistic regression analysis are shown to the right. (C) Posterior predictive checks based on applying a confirmatory hierarchical logistic regression to the original data. Pink bars represent 95% credible prediction intervals. The model captures the original data, but fails to predict the replication's higher performance for competitor pictures from the selective-retrieval stage. (D) Posterior retrodictive checks based on applying a confirmatory hierarchical logistic regression to the replication with priors based on the original data. Grey bars represent 95% credible prediction intervals. The model fails to capture the replication's lower performance with competitor pictures from the baseline stage.**

associated uncertainty (based on the categorical logit model) for the selective retrieval task. Error bars represent the 95% credible intervals generated via simulations from the posterior samples, indicating the range of plausible group means conditioned on the observed data. Similar to our previous attempts, we replicated the qualitative findings of Wimber et al.'s study. On average, subjects correctly picked the category for the target image 71% of the time, and had more intrusions than errors (15% *vs* 9%, with a posterior *p*-value less than .001 for the difference). Again, our subjects had more intrusions and errors relative to Wimber et al.'s participants, but they selected "unknown" less often (5% *vs* 13%).

Panel B presents the trend analysis for hits and intrusions using the mixed effects logistic regression model. The plots on the left show the change in the proportion of hits and intrusions across the 4 selective retrieval trials that tested the same cue word, and the plots on the right show the slope of the linear trend estimated from the mixed effects logistic model. Like the original study and previous replication attempts, subjects once again showed a significant improvement in hits and a decline in intrusions with across the 4 trials.

We evaluated the presence of forgetting effects in the replication data via two methods. First, we fit the confirmatory model to Wimber et al.'s original data, and then conducted a posterior predictive check. If we successfully replicated the original forgetting effect, then the new replication data should fall within the 95% credible intervals based on estimation from the original study. Panel C shows the results of this test. Again, filled circles represent average proportion correct for target pictures, while empty squares represent average proportion correct for competitor pictures. The pink boxes represent the 95% credible intervals for the posterior predictive check for each group mean. The confirmatory model nicely fits Wimber et al.'s original data, but fails to predict the higher average proportion correct for competitor pictures that underwent selective retrieval with the replication data. Note that Wimber et al. had only 24 subjects in their sample compared to the 48 subjects in our pre-registered replication, resulting in more uncertainty in the credible intervals (i.e., wider intervals) and a greater likelihood of finding a successful replication. Despite this bias, we still failed to replicate the original pattern of results.

In our preregistration script, we planned to evaluate the posterior estimates for the effect of forgetting for our confirmatory model. However, this approach is conditional on the confirmatory model successfully fitting the replication data. To evaluate whether the model fit our new data, we conducted a posterior retrodictive check, the results of which are shown in Panel D. Here, the gray boxes represent the 95% credible intervals generated via simulations based on the posterior samples from fitting the replication data, using the posteriors from the previous analysis of the original study as new priors. This approach biases the analysis to favor the results of the original study. Despite this, the confirmatory model fails to fit the replication data, as it cannot capture the lower proportion correct for competitor pictures in the baseline condition. Hence, we failed to replicate Wimber et al.'s results based on two differing tests of replication success. Furthermore, given that the confirmatory model cannot fit our replication data, there is no justification to evaluate the posteriors for the parameters.

Again, applying Wimber et al.'s original analyses to the data for the third replication leads to similar conclusions. The difference between the average proportion of intrusions and errors is statistically significant according to a paired samples $t$-test, $t_{47} = 8.679$, $p < .001$. The one-way repeated measures ANOVA applied to the proportion of intrusions over cue repetitions was significant, $F_{3,141} = 36.5$, $p < .001$, and also exhibited a linear decline, $F_{1,47} = 89.6$, $p < .001$. However, for the final memory task, the repeated measures ANOVA found no significant effects. For the main effect of associate, $F_{1,47} = 2.522$, $p = .119$, while for the main effect of condition, $F_{1,47} = 0.309$, $p = .581$, and finally, for the interaction, $F_{1,47} = 1.909$, $p = .174$. As before, we followed up with the $t$-tests for the planned comparison. The comparison for competitors versus baseline was not significant, $t_{47} = -1.364$, $p = .179$, nor was the comparison for targets versus baseline, $t_{47} = 0.662$, $p = .511$.

We can again examine the significance of the interaction over progressively more powerful statistical tests and differing treatment of missing data. As seen in Table 3, the interaction is not significant regardless of the test. Furthermore, coding missing responses as errors once again had no noticeable impact, since only .3% of the trials were missing.

**Table 3 – The *p*-values for the interaction of associate and retrieval type (Third replication).**

| Test | Scale (Percent) | Missing (error) | Missing (trimmed) |
|---|---|---|---|
| ANOVA | $-\infty$ to $+\infty$ | .174 | .163 |
| Logistic | 0 to 100 | .155 | .157 |
| Logistic (Items) | 0 to 100 | .159 | .159 |
| Multinomial process | 50 to 100 | .106 | .102 |

### 4.4. *Sub-group analyses*

One possible critique of our replication attempts is that our subject composition may have differed substantially from the original study. For instance, astute readers will note that while we replicated the qualitative trends reported by Wimber et al. in the selective-retrieval phase, our subjects on average had more intrusions and errors. A possible concern, then, is that our subjects may not have properly inhibited the competitor images during selective retrieval, which may have reduced the magnitude of any forgetting effects (note that the linear trend analysis weakens this argument, as our subjects clearly produced fewer intrusion with each repetition of the cue word during selective retrieval, indicating that they gradually learned to avoid responding with the competitor category). Another possibility is that our subjects did not properly visualize the images during the cued categorization task (This argument is weakened by the fact that we used the same instructions as Wimber et al. for our second and third replications). Nonetheless, we explored both of these possibilities via a sub-group analysis of our third replication. Our large sample size for the third replication study allowed us to select sub-groups with matching or similar sample sizes to the original study, meaning they still had equivalent power to the original design.

We considered two sub-groups: 1) the subjects who had the lowest number of intrusions during cued categorization (low-intrusion), and 2) the subjects who indicated that they could typically or almost always visualize the image during the cued categorization task (high-visualization). To create our first sub-group, we ranked subjects based on their proportion of intrusions and took the first 24 with the lowest values. For the second sub-group, because we included the end-of-study one-question survey, we were able to assess how well subjects were actually able to visualize the images as instructed. We identified 21 subjects who reported that they typically or almost always were able to visualize the images, and we separated out their data. We also determined that subjects did not have to engage in a high degree of visualization to do well in both the selective retrieval phase and the final memory test. There was only a marginally significant correlation of $R = .28$ ($p = .051$) between the degree of visualization and performance on the selective retrieval stage (i.e., correctly recalling targets), and no significant correlation between degree of visualization and overall performance on the final memory test ($R = .18$; $p = .213$). Another benefit of both subgroups is that they excluded three subjects who had an excessive number of overly fast responses (more than 25% of the trials).

Fig. 6 presents our sub-group analyses compared against Wimber et al.'s original results. We denote the results of the original study in red, the results for the low-intrusion group in blue, and the results for the high-visualization group in purple. Panel A shows the average response proportions (hits, intrusions, errors, and unknowns) and their associated uncertainty (based on the categorical logit model) for selective retrieval performance. Error bars represent the 95% credible

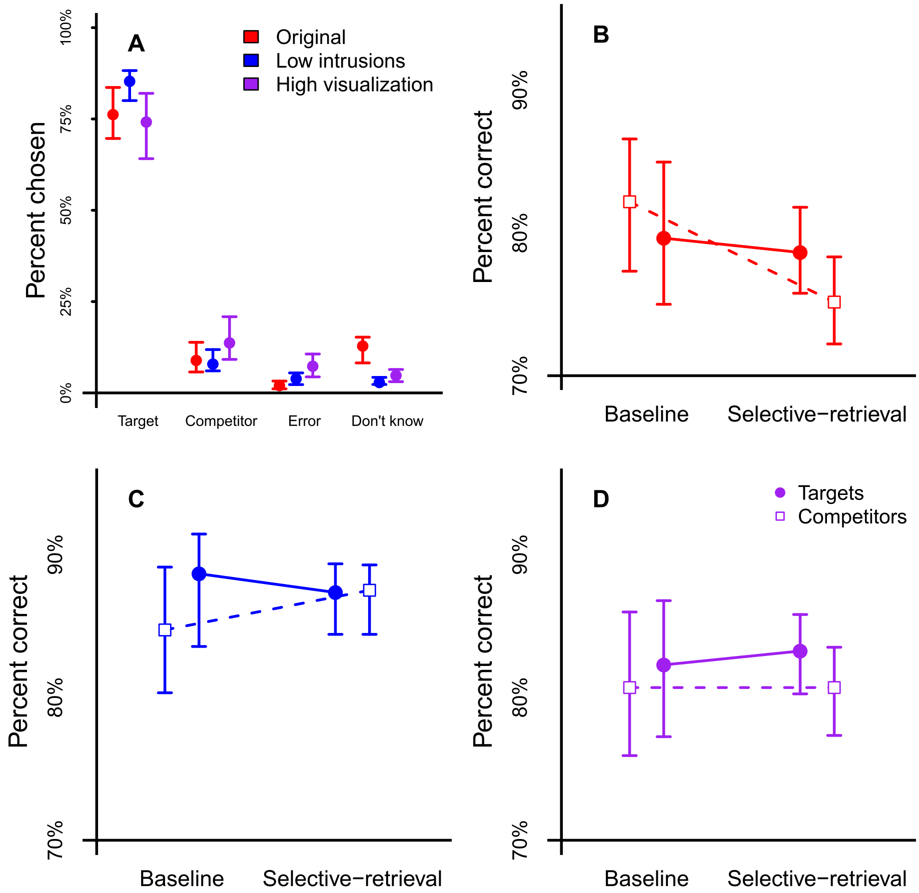

**Fig. 6 – Subgroup analyses of the pre-registered direct replication. The original data is marked in red, the subgroup with a low number of intrusions during the selective-retrieval stage is marked in blue, and the subgroup with high visualization is marked in purple. (A) The average percentage in which the categories for a particular image type (target, competitor, incorrect, or unknown) were chosen collapsed over all selective retrieval trials. Error bars represent 95% credible intervals from a categorical logit model. (B) Performance for the final recognition memory task from the original study with 24 subjects. An interaction is present due to forgetting for competitors but not targets relative to a baseline condition. Error bars represent 95% credible intervals based on a descriptive hierarchical logistic regression model. (C) Performance for the final recognition memory task for the low-intrusions subgroup with 24 subjects. Subjects exhibited learning instead of forgetting for competitors. (D) Performance for the final recognition memory task for the high-visualization subgroup with 21 subjects. Subjects exhibited no forgetting for competitors.**

intervals generated via simulations based on the posterior samples, indicating the range of plausible group means conditioned on the observed data. We can see that the low-intrusion subjects had matching proportions for intrusions and errors relative to the original study (8% *vs* 9%, respectively). Hence, we were able to identify a subgroup with a similar pattern of results for the two most critical response categories compared to the original study. In contrast, the high-visualization group still reported a greater number of intrusions and errors. Note that while subjects in either subgroup for the replications never had as many average "unknown" responses, this is not surprising, as the high average proportion for "unknown" responses in the original study is the result of a single subject who chose the "unknown" option in about 50% of the trials.

Providing additional evidence against the claim that our failure replicate stemmed from too many intrusions during selective retrieval, we next report evidence that the originally reported forgetting effects were primarily driven by subjects who had more intrusions rather than fewer intrusions. Fig. 7 shows a scatter plot in which we predict the magnitude of forgetting for competitors based on the percentage of intrusions subjects experienced during selective retrieval in the original dataset. As can be seen, there is a trend in which subjects with more intrusions had a greater degree of forgetting. This correlation (using Spearman's $\rho$) is significant if one trims out the subject who had 17% missing data (highlighted in red). According to this trend, because our subjects had a greater competitor intrusion rate, we should have observed larger forgetting effects rather than absence of forgetting effects.

Panels B through D present the average proportion correct over the four conditions in the final recognition task for the original study and the low-intrusions and high-visualization sub-groups from the third replication, respectively. Again, filled circles represent average proportion correct for target

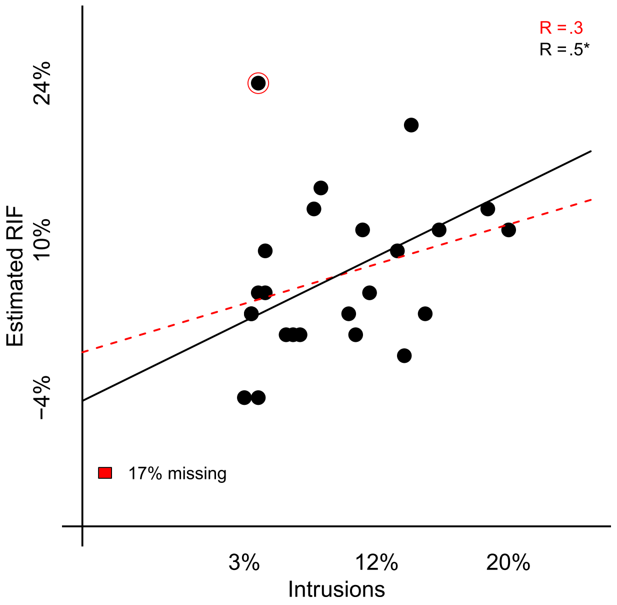

**Fig. 7 – Scatterplot of the percentage of intrusions (incorrectly responding with the category of the competitor picture) during selective retrieval plotted against the magnitude of the estimated RIF effect for the data from the original study. The correlation (using Spearman's $\rho$) and line of best fit are marked in red. If the subject who had 17% missing data is excluded, the correlation is significant, $p$ = .026.**

pictures, and empty squares represent average proportion correct for competitor pictures. Error bars represent 95% credible intervals generated from simulations based on posterior samples. For presentational purposes, the posterior samples were drawn from a descriptive hierarchical logistic regression where each condition mean was assumed to have a separate fixed effects intercept. The core pattern we sought to replicate was the interaction shown in Panel B, with lower performance for competitor pictures that underwent selective retrieval. However, for both subgroups, there was no evidence of forgetting for competitors. In other words, we failed to replicate the pattern of findings reported by Wimber et al. even though we identified a subgroup with similar performance in the selective-retrieval stage and we isolated a subgroup that engaged in a high degree of the visualization emphasized in the instructions.

## 5. Discussion

Across multiple replication attempts, we failed to replicate the finding of forgetting for competitors reported by Wimber et al. for their behavioral results. The third replication attempt was a direct pre-registered replication, including twice as many subjects as the original study. For this third replication attempt, we identified sub-groups with matching characteristics to Wimber et al.'s data, and yet there was no forgetting even for these sub-groups. Furthermore, even within the original dataset, the interaction between practiced (targets) versus unpracticed (competitors) items that did or did not undergo selective retrieval practice was an artifact of coding missing forced choice recognition responses as errors. Thus, within the original study, if cortical pattern suppression caused forgetting of competitors, then it must be equally true that cortical pattern enhancement caused forgetting of targets considering that there was no greater forgetting for competitors as compared to targets (numerically there was a difference, but this difference was not reliable). This is a radically different interpretation of the neural data, suggesting that retrieval practice caused forgetting, regardless of whether this resulted in the neural response becoming more or less similar to that evoked by to-be-remembered picture. However, even this radically different interpretation of the neural data should be taken with a grain of salt considering that our replications failed to find any forgetting effects.

Our repeated failure to find any forgetting effects with this paradigm came as a surprise; RIF is a well-documented effect occurring even if the final test is recognition. After these repeated failures to find forgetting, the preponderance of the evidence suggests null effects in this paradigm. However, because the original forgetting effect (but not the interaction) appears robust over a variety of statistical tests (see Table 4), irrespective of the treatment of missing data, we speculate on the causes of the published forgetting effect, noting that these are necessarily speculations in light of our failure to reliably produce this effect.

### 5.1. *Inflated false positive rate with an unbalanced design and non-equivalent baselines*

One possibility is that the original finding may have been a false-positive owing to sampling error. One could argue that the magnitude of the forgetting effect for competitors contradicts this argument (7% or 5%, when missing data were treated as missing rather than errors). However, the traditional repeated measures statistical test used to assess the reliability of this effect can be misleading in light of the unbalanced design used by Wimber et al. (54 trials for the selective retrieval conditions and only 18 trials for the baseline conditions). More specifically, the repeated measures ANOVA analysis did not "know" that some of the probability values going into the analysis were based on three times fewer data points than other probability values. To make this clear, consider an extreme example in which all of the data are generated from pure chance (coin flips), with 3 of the 4 conditions reflecting the average of 100 such coin flips for each subject, while the fourth is just a single coin flip for each subject. The probability numbers used in the repeated

**Table 4 – The *p*-values for the simple comparison between competitors that did and did not undergo selective retrieval.**

| Test | Scale (Percent) | Missing (error) | Missing (trimmed) |
|---|---|---|---|
| t-test | $-\infty$ to $+\infty$ | < .0001* | .003* |
| Logistic | 0 to 100 | .002* | .016* |
| Logistic (Items) | 0 to 100 | .001* | .012* |
| Multinomial process | 50 to 100 | .002* | .005* |

* Significant at the $\alpha$ = .05 level.

measures analysis for the 3 conditions with 100 coin flips will of course hover around 50%. However, the probabilities for the fourth condition will all be 0% or 100%, and, with a limited sample size (e.g., just 10 subjects), this fourth conditions could easily reveal a performance level that is radically different than chance (and this might be deemed reliable owing to the reliability of the other conditions, which are assumed by the statistical null hypothesis model to have been generated in the same manner).

The unbalanced design used by Wimber et al. was not as extreme as in this example, and yet there is clear evidence of such sampling effects within the original dataset. For instance, consider the 15th subject. This subject had chance performance during initial training (45% correct, with a Bayes factor ratio of 13 in support of the null hypothesis). During the selective retrieval stage, this subject chose the "unknown" category in 50% of the trials and only identified the category for targets 31% of the time during the selective retrieval phase (If a person chose to guess instead of choosing the "unknown" category, chance performance would be 33%). Note that subjects received feedback during this stage and yet this subject did not appear to learn from this feedback, as a trend analysis using logistic regression indicated that the number of times subject 15 chose the target category was unchanged across the four repetitions of the same cue word ($p = .746$). Instead, it appears as if subject 15 became more confused, as there was an increase in the proportion of 'unkown' responses ($p = .009$). In summary, it is clear that this subject failed to encode and recall the target images, and thus his or her final forced choice recognition performance should reflect pure guessing (i.e., coin flips). Nevertheless, subject 15 had an estimated competitor forgetting effect of 11%, which is greater than the average of 7% reported by Wimber et al. If one trims out the missing recognition responses, subject 15 still had an estimated forgetting effect of 7%, which exceeds the average of 5% after trimming missing responses. The point of this example is that the average of 18 coin flips (the baseline condition in the absence of any memory) will be much more variable than 54 coin flips (the competitor condition in the absence of any memory), and this variability can give the appearance of a difference (even though there cannot be a 'real' difference considering that this subject never learned the pictures in the first place).

As this example makes clear, the unbalanced design produces variability unacknowledged by repeated measures ANOVA. This variability becomes particularly pronounced in situations of pure chance (i.e., for subjects who failed to learn the pictures in the first stage of the experiment, meaning that their forced choice responses were necessarily random). This is particularly problematic because the experimental design involves non-equivalent baselines for targets versus competitors, considering that competitors were only studied and tested once during initial learning whereas targets were studied and tested twice. Thus, guessing was likely to play a larger role for the small number of trials used to assess the competitor baseline, and this larger role for guessing would produce greater variability (i.e., an inflated false positive rate for a statistical test that did not include these factors).

We formally addressed the role of guessing via a multinomial process model as applied to this unbalanced design. Besides addressing the number of data points in each condition, the multinomial process model provides a natural way to control for the non-equivalence between targets and competitors owing to different degrees of initial training. We developed a simple null hypothesis model in which initial forced choice performance was a direct indicator of final forced choice performance (remember that the same picture pairs were used during initial learning and during the final recognition test). This is a null hypothesis model because it assumes no effect of selective retrieval practice (neither learning nor forgetting for either targets or competitors). Using this model, we obtained 95% confidence intervals via a bootstrap procedure, specifying the distribution of expected differences from baseline performance. For a given probability of recognition, as indicated by initial performance using a correction for guessing (Equation (1)), the number of pictures correctly identified in the 18 baseline trials and 54 selective retrieval trials was simulated a large number of times, and for each simulation, the difference in percentage correct was computed. Based on 10,000 iterations of this procedure, the 2.5% and 97.5% quantiles specified the outer edge of the gray shaded 95% confidence interval seen in Fig. 8, which plots the relationship between initial performance and final performance: The y-axis indicates final test performance differences from baseline, where positive values are forgetting and negative values are learning. The figure nicely demonstrates that as initial performance approaches the chance level of 50% (i.e., pure guessing), the predicted magnitude of forgetting or learning due to chance alone can vary up to 30 percentage points.

In Fig. 8, we also overlaid each subject's observed degree of forgetting/learning for targets (in black) and competitors (in blue) against their observed level of initial performance. As seen in the figure, nearly all of the observed values fall within the gray shaded 95% null hypothesis region, with the proportion outside of this region commensurate with a 5% false positive rate. This comparison reveals an interesting difference between the original study (Panel A) and the preregistered replication study (Panel B). Note that we trimmed missing data for the results reported from the original study. The subjects from the study of Wimber et al. had lower initial performance levels in general, with this particularly true for the competitors. In other words, the subjects from the original study had notably worse performance for competitors relative to our replication attempt, and this null hypothesis model makes it clear that this is exactly the circumstance in which the unbalanced design can produce spurious results.

### 5.2. *Verbal overshadowing*

In light of the two failures to replicate, and in light of the unacknowledged role of guessing in an unbalanced design with non-equivalent initial training, we conclude that the original forgetting effect in the published study is likely a false positive. One might be tempted, however, to argue that the conclusion of a false positive, even with two failed replications, is premature. After all, the RIF paradigm has been replicated many times, finding robust forgetting effects even with recognition memory (Spitzer, 2014), and even when testing recognition of pictures (Maxcey and Woodman, 2014). From this perspective, the priors for a forgetting effect are

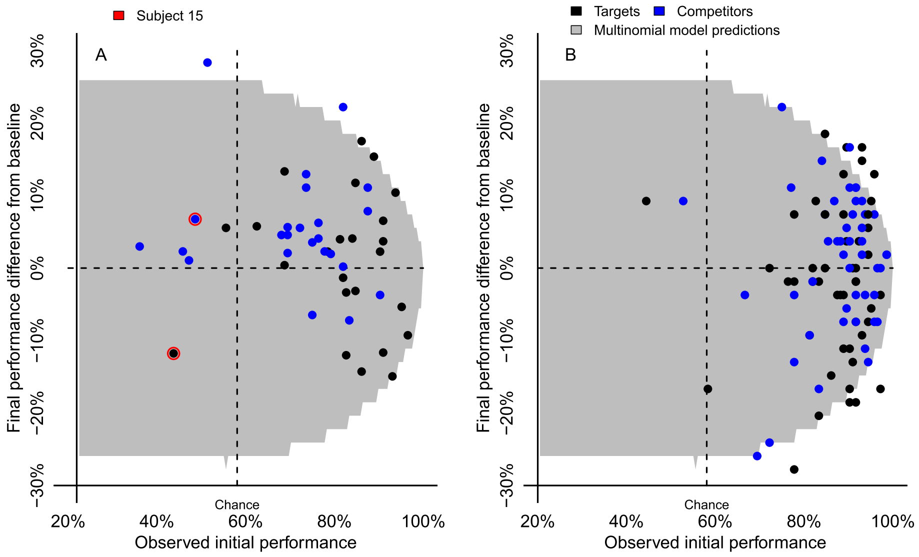

**Fig. 8 – A comparison between observed values and a null hypothesis confidence interval that addresses the role of guessing in the unbalanced design, with three times fewer trials used in the baseline conditions. The *y*-axis shows the difference between final performance after selective retrieval as compared to baseline, with positive numbers indicating forgetting and negative numbers indicating learning. The *x*-axis shows forced choice performance from the initial practice test (the second practice test for targets or the only practice test for competitors). The 95% confidence interval is the gray shaded region, revealing greater variability when initial performance is near the chance level of 50%. This confidence interval assumes no effect of selective retrieval, with final performance only reflecting initial performance and random guessing. Each point shows the observed performance of a single subject for either targets (in black) or competitors (in blue). Panel A shows the results for the original study, while Panel B shows the results for the pre-registered replication. Missing responses were trimmed. As seen in the figure, the proportion of observed values outside of the gray shaded region is commensurate with this being a 95% null hypothesis confidence interval (e.g., approximately 5% fall outside the region, with no greater tendency for this to occur for forgetting as compared to learning). The figure also demonstrates that subjects in the pre-registered replication learned the picture pairs more effectively, as indicated by higher initial performance.**

high (although note that our statistical analyses used the original dataset to set the priors, already biasing the results in favor of a forgetting effect). However, the Wimber et al. study is not like many previous RIF experiments in one critical way; this paradigm involves a mismatch between the modality of the retrieved content during retrieval practice versus the modality of final testing. Although subjects were instructed to visualize the image during selective retrieval, the actual task they completed during this phase was cued categorization (verbal, rather than visual). Irrespective of their visual memory, subjects only had to correctly categorize the target picture as belonging to one of the ‘face’, ‘object’, or ‘scene’ categories. In contrast, the final test was a forced choice between highly similar pictures (i.e., two pictures that had equivalent verbal descriptions). Prior RIF studies with recognition as the final test have used categorical recall during retrieval practice (e.g., Hicks & Starns, 2004), and many RIF studies have used different retrieval practice tasks as compared to the final test (Aguirre, Gómez-Ariza, Andrés, Mazzoni, & Bajo, 2017; Saunders, Fernandes, & Kosnes, 2009; Veling & van Knippenberg, 2004). However, in these studies, the retrieved content during retrieval practice was matched to the final test; if the retrieval practice involved words, the final test was for words, and if the retrieval practice required a response to pictures, then the final test was for pictures. Thus, it may be that the modality mismatch in the Wimber et al. study caused forgetting in a different manner than other RIF tasks.

We reiterate that the most parsimonious explanation is that the original result was a false positive, but we briefly speculate on an alternative explanation (considering that we were unable to replicate the forgetting effect, we were unable to test this alternative explanation).

The act of verbally categorizing a visual memory is thought to underlie forgetting from ‘“verbal overshadowing” (VOS; Schooler & Engstler-Schooler, 1990), and this could potentially explain the original results. The VOS paradigm is remarkably similar to the Wimber et al. paradigm: in VOS paradigms, the first stage presents information visually (e.g., a video of a robbery), the second stage involves verbal categorization of the visually imagined memory (e.g., a verbal description of the robber), and the final stage is forced choice visual recognition between items that have the same verbal description (e.g., an eyewitness lineup). Forgetting effects in VOS are thought to reflect a distortion process in which the visual memory is altered by verbal categorization (e.g., by verbally labeling the race of a criminal, the visual memory of an eyewitness is altered to reflect a more stereotypical version of that race). This distortion process critically needs both visualization (as per instructions in the Wimber et al. study) and a verbal categorization judgment (the overt responses given during selective retrieval). The subjects in the Wimber et al. study were clearly visualizing not only targets but also competitors during selective retrieval as demonstrated both by their erroneous selection of the competitor category and by above baseline pattern similarity for competitors on the first selective retrieval trial. Thus, if the mental images they created were distorted by the verbal categorization task, this would cause forgetting for both targets and competitors.

A recent large replication study of VOS found that it is a reliable effect, and, furthermore, that VOS is more likely to occur for weak visual memories (Alogna et al., 2014). By example, if you see someone in passing, and later report their race, your visual memory of their face may be easily distorted. In contrast, reporting the race of someone in your family is unlikely to alter the strong visual memory you hold for their face. This interaction between VOS and memory strength may explain why the original results revealed numerically greater forgetting for competitors than for targets and it may also explain why we failed to find any forgetting effects. More specifically, because the competitor pictures were only studied once during initiation training, whereas target pictures were studied twice, the visual memories for the competitors was likely weaker than that of targets, making the competitors more susceptible to VOS. On this account, it is not the pattern suppression by the fourth trial of selective retrieval that explains forgetting but rather the pattern enhancement on the first trial of selective retrieval that was the true cause of forgetting (there was a significant pattern enhancement for competitors on the first trial, indicative of competitor intrusions). This account explains why there was a main effect of forgetting, with forgetting arising from pattern enhancement for both targets (by the end of selective retrieval) and competitors (at the start of selective retrieval). This account may also explain our failures to replicate. More specifically, because our subjects were more effective in their initial learning of the pictures (as revealed by performance on the initial forced choice recognition testing), their stronger visual memories may have been shielded from verbal overshadowing distortions during the selective retrieval task.

### 5.3. *Conclusions*

Why does repeated retrieval of a new phone number cause forgetting for an old phone number? Using behavioral measures, retrieval induced forgetting effects are equally well-explained by targeted inhibition (e.g., inhibition of the old phone number) and learning/interference (e.g., competition from the new phone number). The study of Wimber et al. pioneered a method for tracking individual item responses with fMRI pattern analyses to investigate the causes of forgetting. However, a demonstration that inhibition caused forgetting necessarily relies on the behavioral outcome; inhibition cannot be identified as the of the cause of forgetting if there was no forgetting (i.e., if the forgetting effect was a false positive). Furthermore, establishing the cause of something requires a manipulation that produces one outcome in one condition (e.g., forgetting following pattern suppression) but not the other condition (e.g., relatively less forgetting following pattern enhancement), which can be restated as a difference of differences (i.e., a statistical interaction). Our reanalysis of the original behavioral data demonstrated that this interaction was an artifact of labeling missing trials as errors, and thus the original data do not support greater forgetting for competitors as compared to targets. Furthermore, across two replication attempts, we failed to find any forgetting with this experimental paradigm, suggesting that the original forgetting effect was a false positive. In summary, while the neural classifier techniques developed by Wimber et al. will be useful for testing theories of forgetting, the outcome of that study does not provide evidence to favor an inhibition account over other explanations.

## Declarations of interest

None.

## Acknowledgments

This work was supported by the National Science Foundation grant BCS-1431147. We are also indebted to Maria Wimber, who very kindly provided us with her experimental materials, subject data, and analysis scripts.

## Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.cortex.2018.03.026.

## REFERENCES

Aguirre, C., Gómez-Ariza, C. J., Andrés, P., Mazzoni, G., & Bajo, M. T. (2017). Exploring mechanisms of selective directed forgetting. *Frontiers in Psychology, 8*, 1–15. https://doi.org/10.3389/fpsyg.2017.00316.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., et al. (2014). Registered replication report. *Perspectives on Psychological Science, 9*, 556–578. https://doi.org/10.1177/1745691614545653.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology. Learning Memory and Cognition, 20*, 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review, 102*, 68–100. https://doi.org/10.1037/0033-295X.102.1.68.

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomics Bulletin and Review, 11*, 125–130. https://doi.org/10.3758/BF03206471.

Hintzman, D. L. (1986). "schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428. https://doi.org/10.1037/0033-295X.93.4.411.

Jaeger, F. (2007). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007.

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review, 120*, 852–872. https://doi.org/10.1037/a0034246.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron, 93*, 480–490. https://doi.org/10.1016/j.neuron.2016.12.041.

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences, 108*, 5903–5908. https://doi.org/10.1073/pnas.1016939108.

van Maanen, L., Forstmann, B. U., Keuken, M. C., Wagenmakers, E.-J., & Heathcote, A. (2016). The impact of MRI scanner environment on perceptual decision-making. *Behavior Research Methods, 48*, 184–200. https://doi.org/10.3758/s13428-015-0563-6.

Maxcey, A. M., & Woodman, G. F. (2014). Forgetting induced by recognition of visual images. *Visual Cognition, 22*, 789–808. https://doi.org/10.1080/13506285.2014.917134E.

Morey, R., Hoekstra, R., Rouder, J., Lee, M., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*, 103–123. https://doi.org/10.3758/s13423-015-0947-8.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626. https://doi.org/10.1037/0033-295X.89.6.609.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14*, 1105–1107. https://doi.org/10.1038/nn.2886.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110*, 611–646. https://doi.org/10.1037/0033-295X.110.4.611.

R Core Team. (2017). *R: A language and environment for statistical computing*. https://www.R-project.org/version 3.4.1.

Raaijmakers, J. G. W., & Jakab, E. (2013). Is forgetting caused by inhibition? *Current Directions in Psychological Science, 22*, 205–209. https://doi.org/10.1177/0963721412473472.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*, 93–134. https://doi.org/10.1037/0033-295X.88.2.93.

Richards, B. A., & Frankland, P. W. (2017). The persistence and transience of memory. *Neuron, 94*, 1071–1084. https://doi.org/10.1016/j.neuron.2017.04.037.

Saunders, J., Fernandes, M., & Kosnes, L. (2009). Retrieval-induced forgetting and mental imagery. *Memory & Cognition, 37*, 819–828. https://doi.org/10.3758/MC.37.6.819.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36–71. https://doi.org/10.1016/0010-0285(90)90003-M.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/0956797611417632.

Spitzer, B. (2014). Finding retrieval-induced forgetting in recognition tests: A case for baseline memory strength. *Frontiers in Psychology*, 5, 1–6. https://doi.org/10.3389/fpsyg.2014.01102.

Stan Development Team. (2017a). *Bayesian applied regression modeling via Stan*. http: //mc-stan.org. r package version 2.15.3.

Stan Development Team. (2017b). *RStan: the R interface to Stan*. http: //mc-stan.org. r package version 2.16.2.

Stan Development Team. (2017c). *Stan modeling language users guide and reference manual*. http: //mc-stan.org. version 2.16.0.

Tomlinson, T. D., Huber, D. E., Rieth, C. A., & Davelaar, E. J. (2015). An interference account of cue-independent forgetting in the no-think paradigm. *Proceedings of the National Academy of Sciences, 106*, 15588–15593. https://doi.org/10.1073/pnas.0813370106.

Veling, H., & van Knippenberg, A. (2004). Remembering can cause inhibition: Retrieval-induced inhibition as cue independent process. *Journal of Experimental Psychology Learning Memory and Cognition, 30*, 315–318. https://doi.org/10.1037/0278-7393.30.2.315.

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience, 18*, 582–589. https://doi.org/10.1038/nn.3973.