Cognition 211 (2021) 104655

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Refreshing and removing items in working memory: Different approaches to equivalent processes?

Evan N. Lintz [*], Matthew R. Johnson

Department of Psychology, University of Nebraska-Lincoln, United States of America

ARTICLE INFO

Keywords:
Working memory
Refreshing
Removal
Directed forgetting
Directed remembering
Attention

ABSTRACT

Researchers have investigated “refreshing” of items in working memory (WM) as a means of preserving them, while concurrently, other studies have examined “removal” of items from WM that are irrelevant. However, it is unclear whether refreshing and removal in WM truly represent different processes, or if participants, in an effort to avoid the to-be-removed items, simply refresh alternative items. We conducted two experiments to test whether these putative processes can be distinguished from one another. Participants were presented with sets of three words and then cued to either refresh one item or remove two items from WM, followed by a lexical decision probe containing either one of the just-seen words or a non-word. In Experiment 1, all probes were valid and in Experiment 2, probes were occasionally invalid (the probed word was one of the removed/non-refreshed items). In both experiments, participants also received a subsequent surprise long-term memory test. Results from both experiments suggested the expected advantages for refreshed (or non-removed) items in both short-term response time and long-term recognition, but no differences between refresh and remove instructions that would suggest a fundamental difference in processes. Thus, we argue that a functional distinction between refreshing and removal may not be necessary and propose that both of these putative processes could potentially be subsumed under an overarching conceptual perspective based on the flexible reallocation of mental or reflective attention.

1. Introduction

Working memory (WM) is the system that allows us to maintain concurrent active representations of information over a period of time for future access or manipulation in service of goals or actions (Baddeley & Hitch, 1974). Attention is thought to underlie many WM processes, and much research has focused on the mechanisms by which attention operates on WM representations (Cowan et al., 2005; Engle, 2002; Kane, Bleckley, Conway, & Engle, 2001). Many studies have focused on the boundaries of WM function, the cardinal limitation being capacity. Capacity limits vary between individuals and are to some extent contingent upon the properties of the representations being maintained and the current objective. It is generally accepted that WM capacity is confined to roughly three to five items (Cowan, 2010).

Clearly, representations must exit WM at some point. We are obviously able to modify the current WM set, prioritizing the selection of new items when we determine that current items may now be irrelevant. This intuitive and simple concept has spawned a number of lines of research focusing on how WM representations are discarded and the degree to which this can be actively controlled. Some of the first explorations of this yielded an effect known as directed forgetting (DF; MacLeod, 1998; Muther, 1965). The DF field is widely varied, and not all DF studies involve WM manipulations; while aspects of individual designs are contentious, DF paradigms generally present items (individually, as in the item method, or in groups, as in the list method) that are later cued as to-be-remembered (TBR) or to-be-forgotten (TBF). Subsequent tests of long-term memory reveal poorer memory strength for the TBF-cued items than the TBR-cued items, and this effect is ascribed to the intentional forgetting of the TBF-cued information. Though usually interpreted as a forgetting effect on the TBF items, the DF phenomenon could also be interpreted as a consequence of strengthening memory for the TBR items, although most DF studies are not designed in a way that would distinguish between these interpretations.

A related line of work is more specifically focused on similar phenomena at the WM process level, wherein the process is termed WM removal (Lewis-Peacock, Kessler, & Oberauer, 2018; Oberauer, 2001). The research questions are somewhat different than those of DF but the basic paradigms remain quite similar; items or groups are presented and

* Corresponding author at: Department of Psychology, University of Nebraska-Lincoln, Lincoln, NE 68588, United States of America.
E-mail address: evan.lintz@huskers.unl.edu (E.N. Lintz).

https://doi.org/10.1016/j.cognition.2021.104655
Received 29 August 2020; Received in revised form 26 January 2021; Accepted 2 March 2021
Available online 20 March 2021

a subset is cued for removal, while other items must be maintained. When WM is subsequently probed, evidence of successful removal can be inferred in a number of ways; for example, via shorter response times (RTs) suggesting less WM load, decreased set size effects, or decreased classifier evidence for the representations cued for removal (e.g., Lewis-Peacock et al., 2018). Some studies also discuss a putative WM *updating* process in which new items replace older items in the memory set, although one perspective holds that removal may be a component of WM updating responsible for freeing WM capacity in order to accommodate new items (Ecker, Oberauer, & Lewandowsky, 2014).

In each of the preceding examples, and particularly in an updating paradigm, some process of selection is clearly taking place, whether it is selecting certain representations to prioritize over others, selecting representations to discard, or selecting those to retain. Regardless of the mechanism through which these effects are observed, attention is implicated as playing a central role in that selection. Lewis-Peacock et al. (2018) have suggested that being strategy- or goal-oriented is a primary attribute of removal and note that removal may involve withdrawing attention from the remove-cued items, although they were somewhat agnostic regarding the latter point. However, in considering the possibility of an active removal process used to select representations to discard, we face the same paradox raised by studies of thought suppression (e.g., "try not to think of a white bear"; Wegner, Schneider, Carter, & White, 1987). In both cases, it seems intuitive that if information is no longer relevant or desired, one strategy would be to focus elsewhere to avoid it. In the WM context, the most likely targets of one's attention would be the non-removed, or retained, items. This interpretation is consistent with findings that the removal of some items increases the representation strength for the non-removed items and/or makes them more accessible.

Directing mental, or reflective, attention towards an item in WM is often referred to as the cognitive process of "refreshing," which has been found to strengthen a refreshed item's representation relative to the other items in that WM set (Johnson, 1992; Raye, Johnson, Mitchell, Greene, & Johnson, 2007). This foregrounding of the representation increases its accessibility (as observed in RT measures) and leads to improved long-term memory relative to items that were not refreshed. In the lab, refreshing is frequently initiated with a retro-cue indicating which item of a set should be refreshed.[1] Here, we begin to see parallels between refreshing and removal paradigms; items are encoded into WM and a subset are cued as distinct in some way, leading to relative differences in performance. Similar to the way DF results could be alternately framed as effects on the TBF or TBR items, it is possible that the distinction between the effects of refreshing and removal in WM could mainly be one of interpretation. In other words, participants in a WM removal paradigm might engage the strategy of reallocating attention to the remaining non-removed items, i.e., refresh them, or conversely participants in a refresh paradigm might engage the strategy of removing the other items from WM.

In some sense both refreshing and removal occur; it is self-evident that we can explicitly attend to certain items in WM and that representations do not remain in WM forever. But there is still the question of whether the fundamental underlying cognitive processes are the same or different, and whether cognitive models need to include both processes or only one. Theoretical treatments of the subject have, at times, argued for either a refresh-centric or removal-centric perspective of WM that excludes the necessity of the opposite process (Barrouillet, De Paepe, & Langerock, 2012; Lewandowsky & Oberauer, 2015; Lewandowsky, Oberauer, & Brown, 2009; Portrat, Barrouillet, & Camos, 2008), but these arguments generally rely on accepting certain assumptions, or making inferences about WM task data that are open to interpretation; hence the lack of a clear-cut resolution thus far. Certainly, it would help to have empirical data that explicitly pits the processes head-to-head. We are only aware of a single study with an experimental design that directly contrasts these putative processes, which was framed as a study of directed remembering versus directed forgetting in WM (Williams & Woodman, 2012).

In that study, Williams and Woodman used a visual short-term memory task in which three-item subsets of an initial six-item memory set of colored squares were cued either as TBR or TBF. They observed broadly similar results between the two types of cues, with some minor differences, and ultimately did not take a strong stance on whether the underlying mechanisms were the same or different. The use of such a large initial memory set, which is fairly common in studies of DF, updating, and/or removal processes but which exceeds the typical individual's WM capacity, also complicates the theoretical interpretation of such results. Specifically, it becomes somewhat ambiguous to what degree a subject's response to a TBF cue represents "removal" per se, since some items would fail to be maintained even in the absence of a cue. (Consider the metaphor of boxes stacked on a flatbed truck. One can intentionally "remove" a particular box from the truck by picking it up and carrying it away, but intuitively this seems like a much different process than one box falling off the truck because it was overloaded, even if the driver has a certain amount of control over which boxes are more likely to fall off.) Thus, while Williams and Woodman's study provided some amount of insight into the question of refreshing versus removal, more empirical research directly contrasting those processes is still required before the question could be considered resolved.

In the current study, we offer a further attempt to resolve the question of whether refreshing and removal can be experimentally distinguished from one another. Our present design offers several advantages for addressing this question: 1. The WM sets do not exceed three items and thus avoid the complication of set sizes that exceed typical WM capacity. 2. The refreshing and removal versions of our task differ as little as possible, allowing for a direct, low-level comparison between these putatively distinct processes. 3. Measures include both short-term memory probes (RT) and recognition tests of long-term memory (LTM), in order to thoroughly test for potentially different effects of refreshing and removal at different time scales. We also used word stimuli, giving us an additional differentiator from Williams and Woodman's study that could potentially allow us to find differences between processes that might not have shown up using their design.

2. Experiment 1

In this experiment we sought to contrast refreshing and removal by using a paradigm in which the *refresh* and *remove* instructions could be applied as similarly as possible to the same stimuli, with only minor alterations in the refresh/remove cues driving the difference in task instructions. We hypothesized that we would replicate typical refreshing effects (better access to and memory for refreshed items) within the Refresh condition and typical forgetting effects (worse access to and memory for items tagged for removal) within the Remove condition. However, we did not know whether these effects would be equivalent in magnitude (suggesting no practical distinction between refreshing one item vs. removing other items) or exhibit some pattern of differences between effects observed in the Refresh vs. Remove instruction groups (which would be interpreted as evidence for separate refreshing vs. removal processes in WM).

[1] Some researchers have also posited and studied a higher-speed version of refreshing that operates over all WM representations serially, fleetingly, and relatively automatically in service of WM maintenance (Barrouillet et al., 2004; Camos et al., 2018). This "swift" refreshing has been likened to plate spinning, in that each representation is visited in turn, receiving a slight boost in activity that maintains its neural momentum until it is next visited again. In those studies, the tasks are typically more complex, and refreshing is inferred from the data rather than being cued explicitly, but it is otherwise thought to be similar in terms of mechanisms and consequences to the slower and more intentional form of refreshing.

2.1. Method

2.1.1. Participants

Seventy-two undergraduate students (64 female), between the ages of 18 and 35 ($M = 19.63$, $SD = 2.34$), from the University of Nebraska-Lincoln participated in a one-hour experiment for course credit. All participants provided informed consent prior to the study, and procedures were approved by the Institutional Review Board at the University of Nebraska-Lincoln. Even-numbered participants were assigned to the Refresh instruction group and odd-numbered participants were assigned to the Remove instruction group. Participants were not informed that there were different instruction groups until after the experiment had ended; each participant only received instructions for his or her own group (Refresh or Remove). Each pair of participants (01 and 02, 03 and 04, etc.) received identical stimulus sequences on the main task (see below), except for the refresh/remove cues and corresponding instructions. Four participants' data were discarded due to poor compliance with the task directions (e.g., always making the same response) and replaced with subsequently recruited participants.

2.1.2. Procedure

Participants completed each experimental session individually in a private testing room, with all stimuli presented on a standard desktop PC using PsychoPy software (Pierce, 2007). After informed consent was obtained, participants were given instructions for the main task. After the main task instructions, all participants completed a practice session of 12 trials, monitored by the experimenter. Participants who did not understand the task, responded incorrectly, or failed to consistently respond within the allotted time window repeated the practice session. Following the main task, participants completed an unrelated visual short-term memory task, which mainly served as a delay period before the surprise long-term memory (LTM) test that followed. The task lasted approximately ten minutes and followed a standard change-detection task design in which sets of simple colored shapes were presented, followed by a brief delay and then a second shape display in which one of the shapes might have changed color. Participants provided a yes/no response as to whether one of the shapes had changed. This interim task was chosen because it was cognitively demanding, preventing participants from thinking back to the previous task, and because the nonverbal stimuli would not interfere or interact with memories of the main task stimuli. Lastly, participants were informed about and then completed the surprise LTM task.

2.1.3. Main task

At the beginning of each trial (see Fig. 1a), three words were presented simultaneously for 2250 ms. Words were centered horizontally and arranged vertically with the top and bottom items equidistant from the middle item, which was shown at the center of the screen. Next, participants saw a 500 ms blank screen, followed by a retro-cue presented for 1500 ms. The retro-cues consisted of either a single asterisk

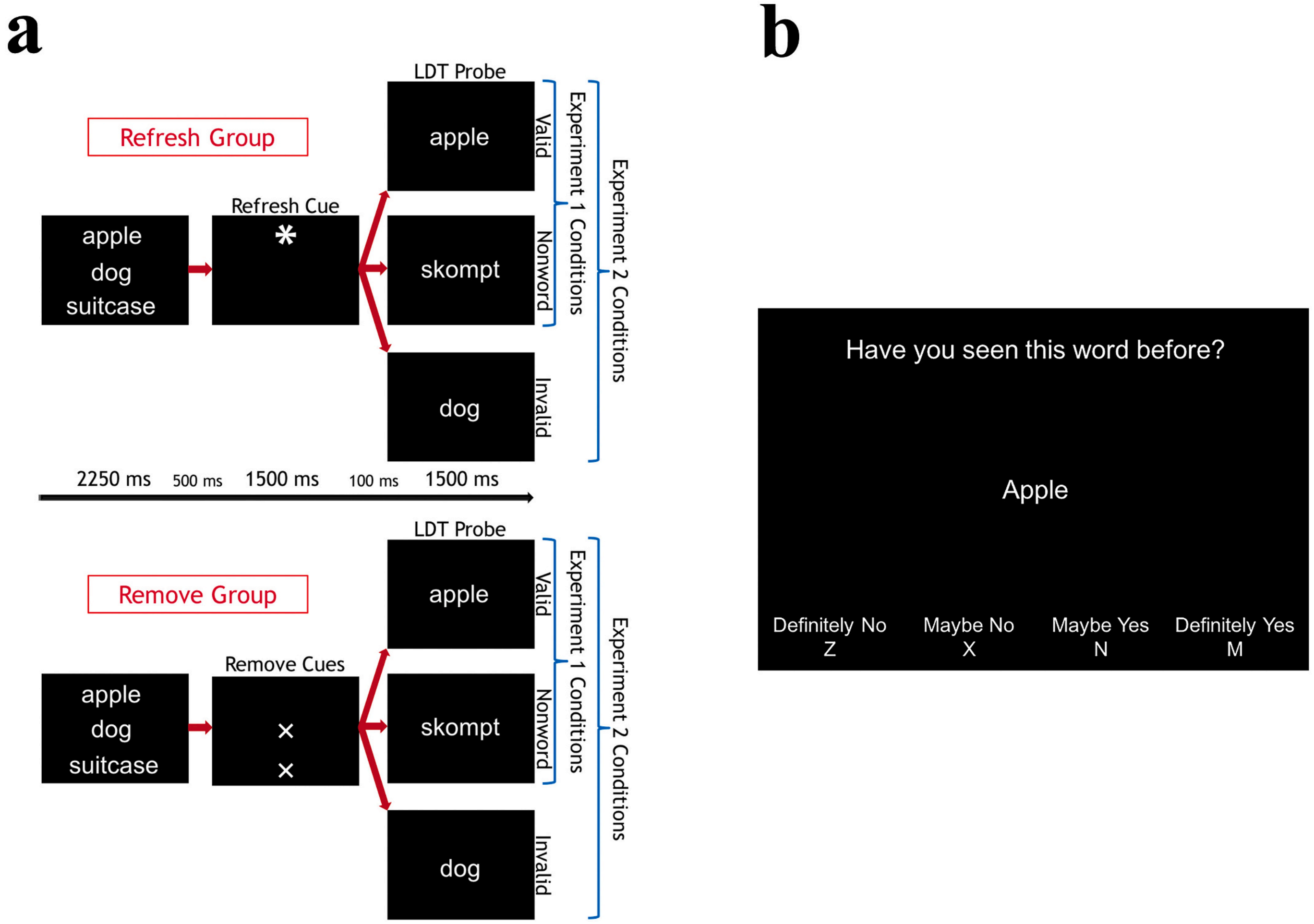

Fig. 1. Task procedures for Refresh and Remove instruction groups in Experiments 1 and 2. a) Main task. Participants saw three words, followed by a cue to refresh one (Refresh group) or remove two (Remove group) item(s) and then a lexical decision task (LDT) probe. The LDT probe could be either the valid item (Experiments 1 and 2), a nonword (Experiments 1 and 2), or an invalid item (Experiment 2 only). b) Example screen from surprise long-term memory (LTM) test that followed the main task. The onscreen appearance of the LTM task was identical across instruction groups, and across Experiments 1 and 2.

shape indicating which word should be refreshed (Refresh instruction group) or two cross (×) shapes indicating which words should be removed (Remove instruction group), presented horizontally centered in the same vertical location as the words they were cueing. The cues for both groups were shown at equivalent sizes to each other and to the original words. Then, another blank screen was presented for 100 ms, followed by a lexical decision task (LDT) probe in which a word or non-word (e.g., "skompt") was presented centrally for 1500 ms. Following the probe, the screen was blank for a three-second inter-trial interval.

For the initial presentation of the three words, participants were instructed that those words would only be on the screen briefly, so they should silently read the words to themselves before they disappeared. Then, during the cue period, participants in the Refresh instruction group were instructed to think back to the word that was just in that position and told that the cue indicated that the word would be relevant to the rest of the task. For the Remove instruction group, participants were instructed to forget the words that were just in the two cued positions, as they were no longer relevant to the rest of the task. For convenience, we will sometimes use "relevant item" to refer both to the word that is cued for refreshing in the Refresh instruction group and also to the word that is *not* cued for removal in the Remove instruction group. Similarly, we will sometimes use "irrelevant items" to refer both to the two words cued for removal in the Remove instruction group and also to the two words *not* cued for refreshing in the Refresh instruction group. Both groups of participants were told that the purpose of the task was to measure response time to the LDT probe and that complying with the cue directions would maximize their performance.

For the LDT probe, participants were told that either a word or a non-word would appear, and that they should respond with a keypress indicating which it was. Participants responded with the index finger of their dominant hand for words and with their middle finger for non-words. Participants were asked to respond as quickly as possible to the LDT probe, but without sacrificing accuracy. The task contained an equal number of word-probe and nonword-probe trials, and on all word-probe trials, all probes were *valid*. In other words, the probe was always the relevant item, and thus the refresh/remove cues were 100% predictive of the probe word on word-probe trials (but see Experiment 2 for a design in which not all cues were valid).

All words were presented in white against a black background in a large, readable font (Arial, 72-pixel height). Words were sourced from the English Lexicon Project (Balota et al., 2007) database and comprised common, everyday nouns with one or two syllables. From this larger set, a custom Matlab script (MathWorks, Natick, MA) implementing a genetic algorithm for optimization (Lintz, Lim, & Johnson, 2020) was used to generate six word lists for the main task (54 words per list) and an additional list of foil words for the LTM test (see below; 216 foil words). The script equated all lists for the words' length, frequency, number of phonemes, number of syllables, and average time needed to read them aloud (all $p > .7$). Non-words were sourced from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002) and were selected based on the following parameters: between three and ten letters, only legal bigrams, and morphologically ambiguous syllables.

From the 324 words used in the main task (six lists × 54 words each), 108 trials were generated (3 words used per trial). If we call the six lists A1–3 and B1–3, word triplets were formed by randomly combining one word from each of lists A1–3 or one word from each of lists B1–3, thus resulting in 54 "A" triplets and 54 "B" triplets. These sets of word triplets were used for all participants (not re-shuffled between participants). Assignments of word roles were fully counterbalanced by word list (two options: list A for word-probe trials and list B for nonword-probe trials, and vice versa), relevant item (three options for which word in each triplet [1, 2, or 3] was the refreshed/non-removed item), and presentation order (six options for assignment of words 1, 2, and 3 from each triplet to the top, middle, and bottom screen positions), for a total of 36 counterbalancing permutations. The same counterbalancing was used for both the Refresh and Remove instruction groups such that Refresh participant 1 and Remove participant 1 were both presented with the same word triplets, at the same screen positions, and used in the same roles within the trial; hence, 36 participants per group.

The 108 trial configurations were divided into thirds to create stimulus lists for three 36-trial blocks (approximately 5.5 min/block at 8.85 s/trial); the trial order was then randomly shuffled within block for each participant, so that all participants saw the same set of words in block 1, block 2, and block 3, but the specific sequence of trials within each block was unique for each participant. This shuffling within block had the constraints that no more than three consecutive trials would share the same relevant item position (top, middle, bottom) or LDT probe type (word or non-word). For trials requiring a non-word probe, the non-words were selected randomly from the overall non-word stimulus pool with the constraint that their length (in letters) matched the length of the relevant (refreshed/non-removed) word on that trial, i.e., the word that would have been the probe if it were a word-probe trial.

Importantly, the counterbalancing and randomization scheme ensured that corresponding participants from the Refresh and Remove groups received nearly identical stimuli, except for the actual refresh/remove cues (e.g., if Refresh participant 1 was cued to refresh the top word on a given trial, Remove participant 1 was cued to remove the middle and bottom words on their corresponding trial) and the following parameters that were randomized separately for each participant: The particular sequence of trials within each block, the nonword stimuli used on each nonword-probe trial, and the order of words on the LTM test (see below).

2.1.4. Long-term memory (LTM) recognition test

The LTM test consisted of all 324 words encountered in the main task (both refreshed and removed words but not nonwords) as well as 216 foil words that were not previously presented. The LTM word list presentation order was randomized for each participant with the constraint that no more than three consecutive trials would be a foil, relevant, or irrelevant word. The question "Did you see this word before?" remained at the top of the screen for the duration of the test, as did the response choices and corresponding keys (see Fig. 1b). Test words were presented in the center of the screen one at a time and remained on the screen until participants responded. Response options were 'Z' for "Definitely Not," 'X' for "Maybe Not" (middle and index fingers of left hand, respectively), 'N' for "Maybe Yes," or 'M' for "Definitely Yes" (index and middle fingers of right hand, respectively). For analysis, these four responses were coded as confidence ratings on a 1–4 scale (with 4 = "Definitely Yes"). In giving instructions for this task, we were explicit that participants should respond as best they could as to whether they remembered seeing the word during the main task, regardless of whether or not we had asked them to forget it (or think back to it, in the case of the Refresh condition).

2.2. Results

Response time (RT) to the main task LDT probe was tested with a 2 (instruction group: Refresh or Remove) × 2 (probe type: valid or nonword) mixed ANOVA (see Fig. 2a). RTs shorter than 200 ms and/or more than three standard deviations away from the participant's mean RT were excluded from the analysis. (Data removal based on *Z*-scoring was done irrespective of probe type and iteratively, with means and Z-scores recalculated if any RTs were trimmed, until no RTs with Z-scores >3 remained.) RTs corresponding to incorrect LDT responses were also excluded. Accuracy was near ceiling (97.96% on average); in total, 5.92% of all trials were excluded from analysis. There was a main effect of probe type, $F(1, 70) = 53.217$, $p < .001$, with faster responses to valid probes ($M = 468.57$ ms, $SD = 110.28$) than to non-word probes ($M = 506.17$ ms, $SD = 102.58$). There was no main effect of instruction group, $F(1, 70) = 0.283$, $p = .596$, and no probe type by group interaction, $F(1, 70) = 1.863$, $p = .177$.

LTM confidence ratings were tested with a 2 (instruction group:

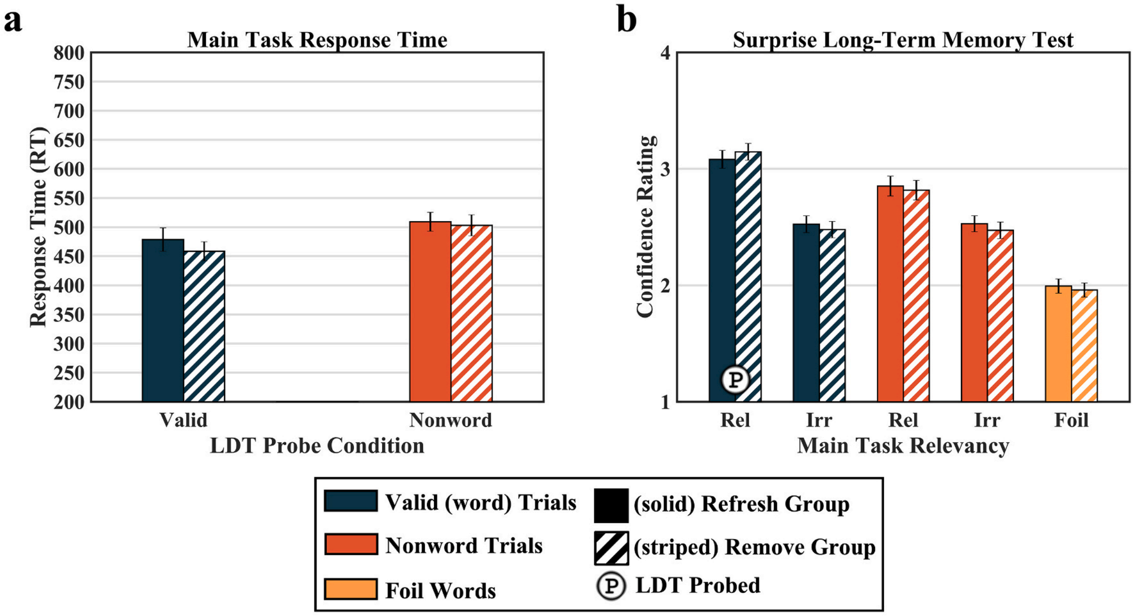

Fig. 2. Results of Experiment 1. a) Main task response time (milliseconds) by instruction group as a function of LDT probe type. Both groups responded faster to valid probes than nonwords, but there was no main effect or interaction according to instruction group (Refresh or Remove). b) Surprise long-term memory test confidence ratings by instruction group as a function of main task word role: Relevant (Rel) or irrelevant (Irr). Bars represent the LTM word's role in the main task (with the exception of foils); for example, the first pair of bars represents the relevant word in a valid-probe trial, whereas the third pair of bars represents the relevant word in a trial with a non-word LDT probe. Both groups showed the expected advantage for relevant (refreshed/non-removed) items over irrelevant ones, but again no main effect or interaction according to instruction group. Error bars indicate the standard error of the mean.

Refresh or Remove) × 5 (word role; see below) mixed ANOVA (see Fig. 2b). These five word roles consisted of foil words plus the four roles that previously-seen words could have played in the main task: relevant words followed by a valid probe, irrelevant words followed by a valid probe, relevant words followed by a non-word probe, and irrelevant words followed by a non-word probe. LTM responses faster than 200 ms were deemed to be accidental keypresses, resulting in 0.23% of total responses excluded from analysis.

For the LTM confidence ratings, there was a main effect of word role, $F(2.149, 150.407) = 352.779$, $p < .001$, Greenhouse-Geisser corrected. There was no main effect of instruction group (Refresh versus Remove), $F(1, 70) = 0.051$, $p = .822$, and no group by word role interaction, $F(2.149, 150.407) = 1.169$, $p = .316$, Greenhouse-Geisser corrected. A planned comparison testing the difference in confidence ratings between relevant ($M = 2.835$, $SD = 0.510$) and irrelevant ($M = 2.501$, $SD = 0.421$) words from nonword-probe trials was significant, $t(71) = 12.899$, $p < .001$, Cohen's $d = 1.520$. The LTM advantage of relevant over irrelevant words is consistent with both traditional directed forgetting and traditional refreshing effects. This comparison based on nonword-probe trials was the most appropriate for assessing the effect of the Refresh/Remove instruction on LTM; valid trials also showed an advantage for relevant over irrelevant words, but that comparison was confounded by the fact that on valid trials, relevant words were seen a second time (as the probe word) but irrelevant words were not. To confirm that both groups showed this advantage and that there was no difference in the magnitude of the advantage between the Refresh and Remove instruction groups, we performed these tests separately in each group. Paired *t*-tests showed a clear advantage of relevant over irrelevant words in both the Refresh ($t(35) = 9.011$, $p < .001$) and Remove ($t(35) = 9.127$, $p < .001$) instruction groups. A two-sample t-test of the difference scores (relevant word ratings – irrelevant word ratings) between instruction groups did not approach significance ($t(70) = 0.412$, $p = .681$).

Additionally, because the previous tests did not support any differences between the Refresh and Remove instruction groups, or any interaction with group, we also performed Bayesian versions of the ANOVAs for main task RT and LTM. One advantage of such Bayesian analyses over traditional null-hypothesis statistical testing (NHST) is that failure to reject the null hypothesis in NHST cannot be straightforwardly interpreted as confirmation of the null hypothesis (i.e., absence of evidence does not imply evidence of absence). However, Bayesian analyses allow us to calculate Bayes Factors (BF) associated with each of several alternative models, which can be expressed either as each model's likelihood relative to a true null model (BF_{10}) or converted to ratios expressing the alternative models' relative likelihood versus each other (Wagenmakers, 2007). Thus, Bayesian analyses can be used to express whether a model with fewer (or no) predictors is actually favored over a more complex one. BFs were computed using the Bayesian ANOVA (BANOVA) functions within JASP (JASP Team, 2018) with default priors (Wagenmakers et al., 2018). The BANOVA for main task RT (see Table 1) found that the best model was a main effect model of probe type (valid or nonword) alone ($BF_{10} = 1.844 \times 10^7$; "extreme" evidence favoring this model over the null). This model was favored over a model that also included a main effect of instruction group, although only in a range considered to represent "anecdotal" evidence against an effect of instruction group (BF = 1.835). Adding a term for an interaction of instruction group and probe type further lowered the

Table 1
Bayesian mixed ANOVA for main task response time (RT), Experiment 1.

Models	P(M)	P(M|data)	BF_M	BF_{10}	Error %
Null model (incl. subject)	0.200	2.974×10^{-8}	1.189×10^{-7}	1.000	
Probe type (PT)	0.200	0.548	4.859	1.844×10^7	8.272
Instruction group (IG)	0.200	1.571×10^{-8}	6.284×10^{-8}	0.528	2.869
PT + IG	0.200	0.299	1.706	1.005×10^7	3.804
PT + IG + (PT × IG)	0.200	0.153	0.720	5.130×10^6	4.433

Note. All models include subject.

likelihood of the model (BF = 1.961) relative to the model including both main effects.

Thus, the Bayesian analyses showed no indication, even at an anecdotal level, that an effect of instruction group or an interaction between group and probe type should be favored over the simpler probe-type-only model. In fact, the simpler model was favored by a factor of 1.8 over a model including an effect of group, and by a factor of 3.595 (1.835 × 1.961) over the model including both main effects and an interaction.

Similarly, a BANOVA of LTM confidence ratings (see Table 2) found that the best model was a main effect model of word role alone ($\mathrm{BF}_{10} = 4.227 \times 10^{104}$; overwhelming support favoring this model over the null). This model was favored over a model that also included a main effect of instruction group (BF = 2.347). Adding a term for an interaction of instruction group and word role further lowered the likelihood of the model (BF = 9.825) relative to the model including both main effects. Thus, in the LTM data, the Bayesian analyses did not favor either an effect of instruction group or an interaction of instruction group and word role, with the simpler word-role-only model being favored by a factor of 2.3 over the model including an effect of group, and by a factor of 23.061 (2.347 × 9.825) over the model including both main effects and an interaction.

2.3. Discussion

For the main task, as was expected, participants were faster overall to respond to words compared to nonwords. There were no differences between the Refresh and Remove instruction groups, which would suggest that both groups were applying a similar strategy to the task. If a fundamental difference between the refresh and removal processes existed, we might have observed various patterns of results that would have been interpretable; for example, the Remove group might have been faster to respond to non-removed words due to a lower overall WM load post-removal, or alternatively, the Refresh group might have been faster to respond to refreshed words due to having their reflective attention more actively placed on the refreshed (non-removed) item. However, we did not observe any such pattern, and our Bayesian analyses were supportive of a null difference between the instruction groups. Due to the strong difference we saw between responses to valid and nonword probes in both groups, it does not seem likely that this lack of difference was due to excessively noisy data or a lack of power to detect reasonably-sized effects. Nevertheless, it remains possible that two of these effects could have canceled each other out. It is also true that in the main task, the only words tested were the relevant words, so it is also possible that any differential effects of the putatively different refresh/removal processes might only have been observable in terms of their effects on the irrelevant items. In order to assess effects on irrelevant items, we considered the LTM analysis from Experiment 1 (see next paragraph) and also ran a second experiment to directly probe effects of refreshing versus removal on RT during the main task (see Experiment 2, below).

Table 2
Bayesian mixed ANOVA for LTM confidence ratings, Experiment 1.

Models	P(M)	P(M|data)	BF_M	BF_{10}	Error %
Null model (incl. subject)	0.200	1.610×10^{-105}	6.440×10^{-105}	1.000	
Word Role (WR)	0.200	0.681	8.522	4.227×10^{104}	0.531
Instruction Group (IG)	0.200	3.475×10^{-106}	1.390×10^{-105}	0.216	1.016
WR + IG	0.200	0.290	1.633	1.801×10^{104}	6.835
WR + IG + (WR × IG)	0.200	0.030	0.122	1.833×10^{103}	14.032

Note. All models include subject.

In addition to yielding insight on irrelevant items, our analysis of the LTM task confidence ratings also affords an alternate way to measure any effects on relevant items that might not have shown up in the RT analyses of the main task. The most pertinent comparison in this analysis focused on possible LTM differences between relevant and irrelevant words on trials where the probe in the main task had been a nonword, which avoids the confound that on valid trials, the relevant word was seen an additional time. In both instruction groups, we found a substantial LTM advantage of relevant over irrelevant words, which would be expected both in traditional directed forgetting paradigms and in traditional refreshing paradigms. However, there was no difference between instruction groups in the size of the "directed forgetting effect" versus the corresponding "refresh effect" (p = .681). Although an advantage of relevant over irrelevant words would be expected in both instruction groups given the corresponding literature in each domain, we would not necessarily have expected the effect to be of exactly the same magnitude. In fact, both groups had nearly identical memory strength for each individual word role in the LTM test; when each word role was independently t-tested between groups, no hint of any potential difference emerged (all $p > .5$). Our interpretation of the NHST results was also bolstered by Bayesian support in favor of a null difference between the instruction groups. Thus, the statistically equivalent pattern of LTM ratings between instruction groups, despite directions to approach the task in very different ways, adds further support to our interpretation that the instructions to refresh versus remove information in WM do not produce measurably different effects on how items are cognitively processed.

3. Experiment 2

In the previous experiment, we found no evidence of any behavioral differences between the Refresh and Remove instruction groups. Thus, Experiment 1 offered no support for, and in fact some support against, a functional distinction between the *refresh* and *remove* processes, although that task design did not allow us to look for certain potential patterns of differences, such as differential effects on irrelevant items during the main task. Specifically, we did not ever probe irrelevant items during the main task, so we could not observe whether there was an RT benefit for relevant over irrelevant items during the initial task, or whether the magnitude of such a benefit might differ between instruction groups. Thus, Experiment 2 was conducted in the same manner as Experiment 1, except for the inclusion of *invalid* probes (i.e., occasional probes of irrelevant words) in the LDT (see Fig. 1a). Adding these invalid probes allowed us to examine the fate of the irrelevant words during the main task and further illuminate any potential differences between instruction groups. We predicted faster RTs to relevant than irrelevant items in both instruction groups, but that any difference between the refresh and removal processes would manifest as an interaction. If there were such a difference between the refresh and removal processes, presumably RTs to irrelevant items in the Remove instruction group would be particularly slowed (versus their simply not being refreshed in the Refresh instruction group), which would drive the hypothesized interaction.

3.1. Method

3.1.1. Participants

Sixty-nine undergraduate students (N = 54 female), between the ages of 18 and 26 (M = 19.74, SD = 1.66), from the University of Nebraska-Lincoln participated in a one-hour experiment for course credit. All participants provided informed consent prior to the study, and procedures were approved by the Institutional Review Board at the University of Nebraska-Lincoln. In Experiment 2, each participant was randomly assigned to either the Refresh or Remove instruction group (N = 30 and 39, respectively). Participants that had already participated in

Experiment 1 were ineligible to take part in Experiment 2. As in Experiment 1, participants were not informed that there were different instruction groups until after the experiment had ended; each participant only received instructions for his or her own group (Refresh or Remove). An additional 18 participants were excluded from the analyses, four due to poor compliance with task instructions (2 Refresh group, 2 Remove group) and 14 due to LDT accuracy below 75% correct (5 Refresh group, 9 Remove group) in any of the three probe type conditions. Exclusions based on accuracy, with the exception of one participant, were due to accuracy specifically on invalid trials. Accuracies for remaining participants were near ceiling (see 3.2. Results).

3.1.2. *Procedure and tasks*

Experiment 1 was fully counterbalanced for all of the key trial variables in order to control for potential confounds, rather than relying on randomization to produce equivalence between conditions. Given the introduction of invalid trials and the resulting increase in the number of permutations of stimuli, the roles and positions of specific words within Experiment 2 were simply randomized rather than counterbalanced. The same constraints on trial presentation remained; no more than three consecutive trials having the same cue position or probe type (valid, invalid, or non-word). Participant instructions were changed slightly to accommodate the invalid trials. Other researchers (Lewis-Peacock et al., 2018) have noted that the presence of invalid probes may complicate the study of WM removal, as invalid trials might negate any incentive to remove. In order to address the presence of invalid probes and incentivize participants to comply with removal cues, we acknowledged that we would rarely probe the irrelevant words; however, since we were measuring reaction time, we informed participants that they would still be faster overall if they followed the cue instructions. Thus, as it was framed to participants, in service of the overall goal of low average reaction times, it was most effective to remove as cued, given the higher percentage of valid and nonword probe trials. In any case, if the existence of invalid trials indeed de-incentivized participants from complying with removal cues, that should become apparent in the data as a reduction in LTM differences between relevant and irrelevant items. As we did ultimately find a preserved LTM benefit for relevant items (see Results), in this case we do not find reason to suspect that the presence of invalid trials presented an issue for participants' compliance with the instructions.

In the first experiment, LDT word probes had an equal probability of being the relevant word or a non-word; the irrelevant words were never tested. The probability of encountering each probe type in the second experiment was 40% non-word, 40% valid, 20% invalid. While this would change the overall ratio of word to non-word probes from equal (as in Experiment 1) to 60% word, we felt that this change would be unnoticeable to participants. This afforded us the opportunity to test a sufficient number of invalid trials while keeping the proportion of invalid to valid trials low enough that invalid probes would not influence participants' strategy. Total trials were decreased from 108 (Experiment 1) to 90 for Experiment 2. Accordingly, the LTM task in Experiment 2 consisted of all 270 words encountered in the main task and 180 foils, keeping the ratio of previously seen words to foils the same as in Experiment 1.

3.2. *Results*

Response time to the main task LDT probe was tested with a 2 (instruction group: Refresh or Remove) × 3 (probe type: valid, invalid, non-word) mixed ANOVA (see Fig. 3a). RTs shorter than 200 ms and/or more than three standard deviations away from the participant's mean RT were excluded from the analysis, as in Experiment 1. RTs corresponding to incorrect LDT responses were also excluded. In total, 6.02% of all trials from analyzed participants were excluded from analysis. As noted above, participants with <75% accuracy in one or more probe type conditions were excluded from analysis; accuracy was high in the remaining participants (95.53% on average). There was a main effect of probe type, $F(1.828, 122.465) = 102.583, p < .001$, Greenhouse-Geisser corrected, with faster responses to valid probes ($M = 548.379$, $SD = 116.661$) than to non-word probes ($M = 663.781$, $SD = 135.021$), $p < .001$, and invalid probes ($M = 679.576$, $SD = 124.733$), $p < .001$. Responses to non-word probes were not significantly different than to invalid probes, $p = .300$. The main effect of instruction group was not significant, $F(1, 67) = 3.727, p = .058$ (Remove group: $M = 660.763$, SD

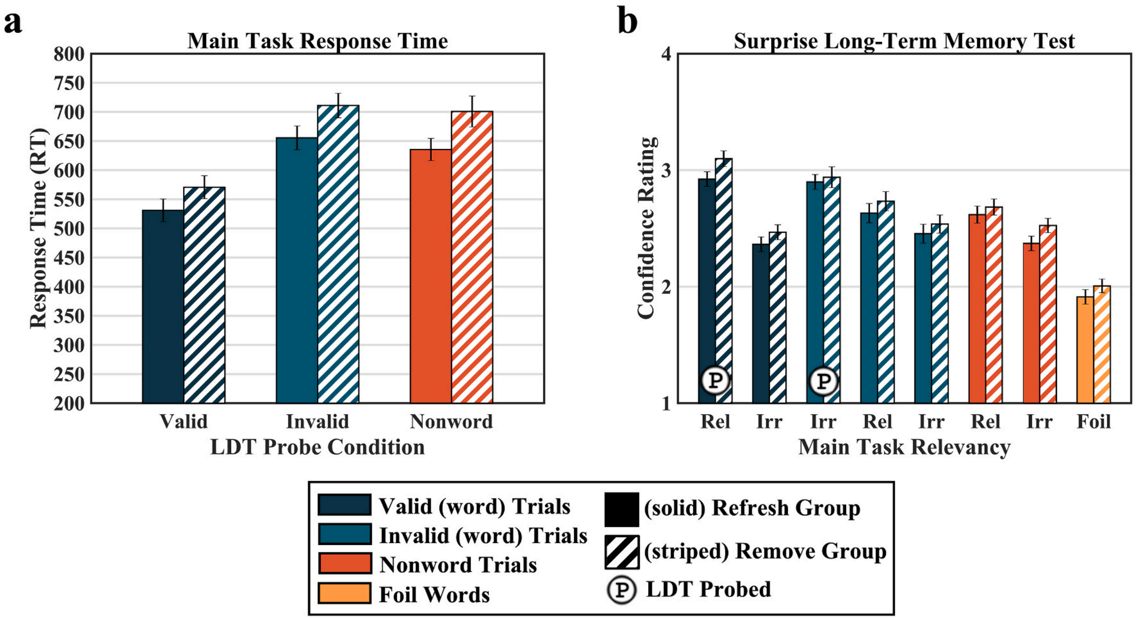

Fig. 3. Results of Experiment 2. a) Main task response time (milliseconds) by instruction group as a function of LDT probe type. Both groups responded faster to valid probes than invalid probes or nonwords, but there was no interaction according to instruction group (Refresh or Remove). b) Surprise long-term memory test confidence ratings by instruction group as a function of main task word role: Relevant (Rel) or irrelevant (Irr). Both groups showed the expected advantage for relevant (refreshed/non-removed) items over irrelevant ones, but no main effect or interaction according to instruction group. Error bars indicate the standard error of the mean.

= 111.773; Refresh group: $M = 607.359$, $SD = 115.525$). There was no significant probe type by instruction group interaction, $F(1.829, 122.465) = 0.821$, $p = .433$.

LTM confidence ratings were tested with a 2 (instruction group: Refresh or Remove) × 8 (word role) mixed ANOVA (see Fig. 3b). Word roles were the same as in Experiment 1, with the addition of three new roles corresponding to the words that had been seen on invalid probe trials (wherein the probe was one of the irrelevant words; thus, the new word roles for invalid probe trials were the relevant word, the irrelevant word that was probed, and the irrelevant word that was not probed). LTM responses faster than 200 ms were deemed to be accidental keypresses, resulting in 0.03% of total responses excluded from analysis.

For the LTM confidence ratings, there was a main effect of word role, $F(4.430, 296.799) = 118.537$, $p < .001$, Greenhouse-Geisser corrected. There was no main effect of instruction group (Refresh versus Remove), $F(1, 67) = 1.454$, $p = .232$, and no group by word role interaction, $F(4.430, 296.799) = 0.538$, $p = .726$, Greenhouse-Geisser corrected. A planned comparison testing the difference in confidence ratings between relevant ($M = 2.651$, $SD = 0.420$) and irrelevant ($M = 2.448$, $SD = 0.365$) words from nonword-probe trials was significant, $t(68) = 6.271$, $p < .001$, Cohen's $d = 0.755$. This replicated the pattern seen in Experiment 1, and again was consistent with both traditional directed forgetting and traditional refreshing LTM effects. No attenuation of the effects was observed in this experiment, suggesting that the inclusion of invalid probe trials in the main task did not generally cause noncompliance with the remove cue instructions. The addition of invalid probe trials in Experiment 2 allowed a second test of the relevant versus irrelevant words that were not subsequently probed, which revealed that relevant, unprobed words from invalid-probe trials ($M = 2.682$, $SD = 0.485$) were rated higher than irrelevant, unprobed words from invalid-probe trials ($M = 2.496$, $SD = 0.467$), $t(68) = 3.959$, $p = .005$, Cohen's $d = 0.477$, thus paralleling the effects seen on nonword-probe trials.

As in the previous experiment, further interpretation of the results was explored with Bayesian analyses. BANOVAs of main task RT and LTM both had broadly similar patterns of results to Experiment 1. For main task RT (see Table 3), the best model did include both a main effect of probe type (valid, invalid, nonword) and a main effect of instruction group ($BF_{10} = 4.422 \times 10^{24}$; extreme evidence favoring this model over the null), although this model's likelihood was only slightly higher than the model including only a main effect of probe type (BF = 1.455). This is consistent with the non-significant trend for a main effect of instruction group ($p = .058$) we saw using NHST; thus, we cannot rule out the possibility of a main effect of group in Experiment 2, but the evidence is not conclusive either way. However, the model including an interaction of instruction group and probe type (as well as both main effects) had lower likelihood than either the model including both main effects (BF = 6.429) or the model including only a main effect of probe type (BF = 4.418). Thus, the magnitude of the evidence against an interaction between probe type and instruction group was similar to that found in Experiment 1.

An LTM task BANOVA (see Table 4) found that the best model was a main effect model of word role alone ($BF_{10} = 5.668 \times 10^{97}$; overwhelming support versus the null model), and including instruction group in the model penalized it by a factor of BF = 1.730. Adding another term for an interaction of instruction group and word role further lowered the likelihood of the model (BF = 93.808) relative to the model including both main effects. Thus, the simplest word-role-only model was favored by a factor of 1.7 over the model including an effect of group, and by a factor of 162.288 (1.730 × 93.808) over the model that also included an interaction.

Thus, the Bayesian analyses suggest that the lack of a significant interaction found with NHST should be interpreted as extremely strong support against the existence of such an interaction.

A final pair of mixed ANOVAs (main task RT, LTM confidence) were performed to directly compare the results of Experiments 1 and 2. For these analyses, only valid-probe and nonword-probe trials were considered, as there were no invalid-probe trials in Experiment 1. For main task RT, the three-way interaction between probe type, Refresh/Remove instruction group, and experiment was non-significant, $F(1, 137) = 0.320$, $p = .572$. For LTM ratings, the three-way interaction of word role, Refresh/Remove instruction group, and experiment was also non-significant, $F(2.472, 338.635) = 0.380$, $p = .728$, Greenhouse-Geisser corrected. Taken together, these results suggest that none of the changes between Experiments 1 and 2 substantively affected the results for any of the conditions shared between the two experiments, and specifically that the inclusion of invalid probes in Experiment 2 did not alter the way in which participants approached the task.

3.3. Discussion

In the main task, as in Experiment 1, participants were faster to respond to valid probes than to nonword probes. Response times to invalid probes were slower than to valid probes, reflecting the expected directed forgetting effect (Remove instruction group) or refresh advantage (Refresh instruction group). Response times were slower overall in Experiment 2 than Experiment 1 (main effect of experiment: $p < .001$), suggesting that the inclusion of invalid probes added some degree of difficulty or ambiguity to the task overall. This is supported by the fact that 13 participants had to be excluded in Experiment 2 due to low accuracy (compared to none in Experiment 1), primarily due to inaccurately responding "nonword" to invalid probes. Also, anecdotally, several participants voluntarily mentioned an impulse to sometimes respond to invalid probes as nonwords, suggesting that the slower overall RTs in Experiment 2 may have resulted from a recognition of this impulse and a desire to limit the number of such incorrect responses. This tendency to occasionally misinterpret the instructions and respond "nonword" to invalid probes may have been slightly stronger in the Remove instruction group, which could account for the non-significant trend towards a main effect of instruction group in Experiment 2.

Table 3
Bayesian mixed ANOVA for main task response time (RT), Experiment 2.

Models	P(M)	P(M|data)	BF_M	BF_{10}	Error %
Null model (incl. subject)	0.200	1.227×10^{-25}	4.909×10^{-25}	1.000	
Probe Type (PT)	0.200	0.373	2.379	3.039×10^{24}	1.443
Instruction Group (IG)	0.200	1.578×10^{-25}	6.310×10^{-25}	1.285	0.738
PT + IG	0.200	0.543	4.747	4.422×10^{24}	4.588
PT + IG + (PT × IG)	0.200	0.084	0.369	6.878×10^{23}	2.062

Note. All models include subject.

Table 4
Bayesian mixed ANOVA for LTM confidence ratings, Experiment 2.

Models	P(M)	P(M|data)	BF_M	BF_{10}	Error %
Null model (incl. subject)	0.200	1.114×10^{-98}	4.455×10^{-98}	1.000	
Word Role (WR)	0.200	0.631	6.848	5.668×10^{97}	0.332
Instruction Group (IG)	0.200	4.475×10^{-99}	1.790×10^{-98}	0.402	0.919
WR + IG	0.200	0.365	2.298	3.276×10^{97}	1.083
WR + IG + (WR × IG)	0.200	0.004	0.016	3.493×10^{95}	1.264

Note. All models include subject.

However, even if true, we would interpret that tendency as an overarching strategic or task-level difference in how the groups interpreted the slightly different task instructions, rather than a process- or item-level difference between refreshing and removal, as we did not see any interaction between instruction group and probe type.

In terms of LTM recognition, Experiment 2 replicated the findings of the first experiment for both the Refresh and Remove instruction groups, for what would generally be referred to as refreshing advantage or the directed forgetting effect, respectively. In both cases, LTM confidence ratings for irrelevant words were lower than for relevant. This was true for words that had occurred in both invalid- and nonword-probe trials, but had not occurred as probes themselves, in the main task. As in Experiment 1, there was no difference between instruction groups in the size of the "refresh effect" versus the "directed forgetting effect" for either probe type ($p = .190$ for words seen on nonword-probe trials; $p = .837$ for words seen on invalid-probe trials). Again mirroring the first experiment, there was no significant difference in memory strength between groups for any individual word role (all individual $p > .06$) and no hint of an interaction between group and word role ($p = .726$). Our interpretation of the NHST results, namely that there was no appreciable difference between the instructions to refresh versus remove information in WM, was again supported by Bayesian analyses favoring no interaction between instruction group and word role in either main task RT or LTM confidence ratings.

4. General discussion

The current study examined WM refreshing and removal tasks with parallel structures in order to determine whether these putatively distinct processes were indeed dissociable in a carefully equated, head-to-head comparison. In the experiments described above, the Refresh and Remove instruction groups received nearly identical word stimuli, differing only in whether the relevant or irrelevant items were cued. Participants were either cued on relevant items and instructed to refresh them or cued on irrelevant items and instructed to remove them, with a task design similar to those commonly employed in both refreshing and removal paradigms. For participants in the Refresh instruction groups in both experiments, results from both of our primary measures (RT to LDT probes in the main task; confidence ratings in a later surprise LTM test) replicated the findings of the previous refreshing and retro-cue literature: Refreshing strengthened the WM trace for that item, enhancing short-term accessibility (faster RT) and leading to improved LTM, relative to the non-refreshed items. Conversely, for participants in the Remove instruction groups in both experiments, results from both primary measures replicated the findings of previous removal and directed forgetting studies: Lower accessibility and worse memory for items cued for removal, compared to non-removed items. The addition of invalid probes in Experiment 2 did slow RTs to LDT probes compared to Experiment 1, and there was a non-significant hint that this may have been slightly more pronounced in the Remove group; however, there was still no evidence of any interactions between word roles and instruction groups, which would be required to demonstrate any difference between the consequences of the Refresh versus Remove instructions at the process or item level. In fact, our Bayesian analyses indicated that there was more evidence *against* models including such an interaction than evidence *for* models including it. Thus, our results seem to be consistent with the possibility that there is no consequential difference between the postulated "refreshing" and "removal" processes at a fundamental level.

Proponents of a distinct removal process might attribute the observed pattern of results for the Remove instruction group (namely, a benefit to the non-removed items) to the abatement of interference or cognitive load that would otherwise be incurred by the continued maintenance of extraneous items. Thus, results would generally be framed in terms of suppressive or inhibitory processes and with a theoretical focus on the items targeted by those processes. From this perspective, the effects observed for the Refresh instruction group (a benefit for refreshed items) could be framed in terms of those suppressive or inhibitory processes as well; other researchers have previously suggested that the cueing of relevant items implicitly indicates the irrelevance of the uncued items, thereby invoking the putative removal processes for those uncued items (Souza & Oberauer, 2016; see also Lewis-Peacock et al., 2018).

On the other hand, refreshing-based interpretations are usually framed in terms of the positive effects of refreshing relative to the non-refreshed items; researchers more focused on refreshing might attribute the pattern of results for the Refresh instruction group to the attentional foregrounding of the relevant items, thus strengthening those representations in WM. To our knowledge, researchers focused on refreshing have not explicitly advocated that "removal" of information is executed by means of refreshing items not cued for removal; theoretical discussions of refreshing typically make few or no assumptions about a cost to non-refreshed items, although some might contend that the non-refreshed items would be more susceptible to interference or time-based decay. However, one could readily extrapolate the inverse point to that raised in the previous paragraph: That, from a refresh-based perspective, cueing some items for removal might engender refreshing effects, as attention is implicitly pushed towards the remaining item(s).

As such, depending on one's theoretical orientation, either refreshing or removal could be re-framed as a roundabout form of the other; in the current study, we found no significant differences in the patterns of either the LDT probe RTs or the LTM confidence ratings contingent upon whether the main task was framed as a refreshing or removal paradigm. This suggests that two distinct, and yet parallel, frameworks for refreshing and removal processes in WM may not be strictly necessary based on the available experimental evidence to date.

Of course, it is still possible that refreshing and removal are two distinct processes, each operating on the inverse items of the other, but with the same apparent objective: to selectively promote the maintenance of a subset of WM representations. However, for this scenario to be consistent with our current observations, either the measurable effects of refreshing and removal would have to mirror each other with almost exactly the same magnitudes (which would seem to be a fairly unlikely coincidence), or our experimental design must have been in some way insensitive to dissociating these processes. Of course, it is certainly imaginable that future experiments using different stimuli or procedures could go beyond the evidence we have offered and find a definitive distinction between refreshing and removal; however, for the time being, we do not suspect that is likely if those future experiments are also adequately balanced and controlled for any confounding factors in the experimental design or task instructions.

Now, one would still expect that at *some* level, there must be some difference between the Refresh and Remove tasks, even if it is simply due to interpreting different sets of physical cues; however, such differences need not imply a fundamental distinction in the cognitive process(es) at the heart of each task. Here, it is important to make a distinction that is sometimes neglected in theoretical treatments of this topic, between *tasks* and the *strategies* used to perform them, versus the core cognitive *processes* that underlie them. For example, we could imagine that one could perform a Refresh task by essentially removing the non-refresh-cued items, or perform a Remove task by refreshing the non-removed items; in either case, one could argue that only a single core WM process is necessary to perform either task. This would still entail an additional minor operation of cue translation/inversion to be inserted before the primary process is invoked, but that operation might be relatively trivial in terms of time and cognitive effort, and might not be detectable with standard cognitive psychology methods. These more subtle differences could potentially be revealed in future studies using more sensitive methodologies such as electroencephalography (EEG; cf. Johnson, McCarthy, Muller, Brudner, & Johnson, 2015).

Similarly, detectable differences could also exist in global strategies people employ in performing a task, without necessarily implying a difference in the core cognitive processes involved. In the main task of Experiment 2, we did observe a trend towards a significant main effect of instruction group, with the Remove group responding slower overall. The Remove instruction group also had almost twice the rate of attrition of the Refresh instruction group, due mainly to high rates of responding "nonword" to invalid-word probes in some subjects. One might hypothesize that the more negatively framed language of the Remove instructions set up an association with Remove-cued items that, in those subjects, conflicted with the ultimate need to answer with a positive "word" response on invalid trials. For the remaining Remove instruction group participants that were able to maintain sufficient accuracy, the slightly slower RT overall may indicate the cost of being more cautious in responding to the probes in order to avoid such mistakes. However, we would still view these differences as reflective of an overall task-level strategy or bias, resulting from the valence of the language in the instructions, but *not* as reflecting a fundamental process-level difference in terms of the (putative) *refresh* and *remove* operations or their downstream effects on item representations.

4.1. Arguments for refresh- versus removal-based accounts

When refreshing and removal processes have previously crossed paths in the literature, researchers have typically justified their case for a single- or dual-process interpretation based on the framework of an encompassing theory (e.g., Morey & Cowan, 2018; Souza & Oberauer, 2016; Souza & Vergauwe, 2018). Given the flexibility of the definitions of these processes, the different measures with which they have been studied, and the clear potential for conceptual overlap between them, it has been difficult to reach a consensus on purely theoretical grounds, or in empirical investigations that primarily seek to confirm one perspective; thus, we believe direct, explicit empirical comparisons of these processes, including the present study, are necessary. We know of only one previous behavioral study, described in the Introduction (Williams & Woodman, 2012) that had a similar design; while that study shed some light on the topic, it did not take a strong stance on this matter in the end. In a computational simulation study, Shepherdson and Oberauer (2018) interpreted their findings as implying that either strengthening and pruning in WM, in their terms (roughly equivalent to our constructs of refreshing and removal, respectively) might not be dissociable at all, or if they were dissociable, a pruning-only mechanism was favored. Now that we have directly compared the two processes behaviorally, we also question whether it is necessary to maintain two separate constructs that seem to achieve the same end state, although we believe there are valid reasons to endorse a refresh-centric model as well.

While we acknowledge that additional empirical investigations will likely be necessary before the issue is settled, there remain reasons we find a refresh-centric perspective compelling. For one, even if the underlying mechanism is equivalent between "refreshing" and "removal" in WM, and although those terms may be equally apt descriptors in certain designs such as our two experiments here (where the refreshing and removal conditions were explicitly set up as precise converses of each other), the refreshing-centric account is both more conceptually and linguistically parsimonious in some contexts. A number of studies have examined refreshing in non-competitive or non-selective paradigms, such as presenting and immediately refreshing a single item (e.g., Johnson, Reeder, Raye, & Mitchell, 2002; Raye, Johnson, Mitchell, Reeder, & Greene, 2002), where the task does not require any further maintenance of the item in WM after the initial act of refreshing or a comparison action (e.g., reading a novel word, pressing a button in response to an abstract cue). One could still contend that those non-refreshing control conditions entail "removing" the initial item from WM in some form, but this phrasing implies a more active form of removal than what seems to take place in those conditions. Conversely, if we consider "refreshing" a solitary item from a removal-only perspective, one would have to re-frame that act of refreshing as a lack of removal, which seems unnecessarily convoluted; it also seems to imply that a condition presented as, and which subjectively feels like, a fairly active mental process is more passive than it actually is.

Despite our relative fondness for a refreshing-centric account of removal (as opposed to the converse, a removal-centric account of refreshing), there remain certain aspects of the pro-removal perspective that are difficult to ignore. For example, some neuroimaging studies have offered decreased classifier evidence for irrelevant items as indicative of a removal process (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012). Even we (Johnson & Johnson, 2009) have found that refreshing a WM representation not only enhanced neural activation for that item, but seemed to suppress the activation of other items, even below their level of activation in a baseline condition in which neither item was explicitly refreshed (or removed). On the one hand, the studies that have most directly compared refreshing and removal do not seem to strongly support the need for two separate fundamental processes; on the other hand, each is conceptually appealing in its own way. Is there another way of conceptualizing a single fundamental process that can subsume both accounts across a wide range of contexts and paradigms?

4.2. The case for a compromise

Specifically, perhaps it would be preferable simply to re-frame both refreshing and removal as complementary aspects of an overarching process of *reallocation* of a finite resource, in the form of mental or reflective attention (Chun & Johnson, 2011; Raye et al., 2007). An analogy we sometimes employ is that attention may be like water, and WM representations like buckets. Suppose we start off with all our water (attention) distributed equally between all the buckets (items), but then decide to pour all the water into a single bucket. Have we strengthened or "refreshed" the contents of the now-full bucket, or have we "removed" those of the now-empty buckets? With the water analogy, these descriptions are complementary but functionally equivalent, and it may be reasonable to contend that the same holds for attention in WM. When a subset of the items currently held in WM are identified as more relevant than others, attention is shifted towards the relevant item(s) in an effort to strengthen their representations and prolong successful maintenance. As a result, attention is likewise withdrawn from other items. This account does not entirely obviate or dismiss the traditional refreshing/removal terminology; both would still be relevant and appropriate descriptors for certain task agendas and behavioral phenomena, just as pouring water from one bucket to another does not mean that it is inaccurate to state that the water was poured *out of* bucket A, or that it was poured *into* bucket B. The different descriptions are more a factor of which bucket's role we want to emphasize in a given scenario.

In this conceptualization, refreshing and removal still occur as *tasks* people perform or *strategies* used to perform them, but the primary core *process* underlying them both is the same. Minor differences could exist – such as the differences in cue translation between a single "refresh" cue or a dual "remove" cue in the present study – but that would more reflect a difference in how the process is initiated, not the process itself. We are certainly not the first to advocate a view of mental or reflective attention like this; attention is already implicated as a central component of several highly regarded theories of WM (Barrouillet, Bernardin, & Camos, 2004; Cowan, 1988; Farrell & Lewandowsky, 2002; Oberauer, 2002). Here, we are simply suggesting that viewing the reallocation of attention as a primary WM process is potentially compatible with the aforementioned models, after a certain amount of reframing. Such a reformulation would alleviate the need for distinct "refresh" and "removal" processes in theoretical formalizations, although those labels might still be used more informally to refer to certain kinds of tasks or behaviors.

We should also recognize that the debate between refreshing and removal resides within a larger disagreement in the literature as to whether a failure to maintain WM representations is mainly due to time-based decay or to interference. As supported so far by the data in the present study, we believe our proposed reallocation account could be viewed as compatible with both decay- and interference-based models and would not necessitate championing one over the other at this time. For example, in one interference-based model, the serial order in a box, complex-span (SOB-CS) model (Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012; see also Farrell & Lewandowsky, 2002), the removal process is presumed to eliminate interference from goal-irrelevant WM items, which might otherwise interfere with the goal-relevant ones. Here, we might frame the SOB-CS "removal" process as a goal-relevant reallocation of attentional resources from irrelevant WM items (i.e., sources of interference), to task-relevant ones. Conversely, in one decay-based model, the time-based resource-sharing model (TBRS; Barrouillet & Camos, 2007), the process of refreshing operates quickly and serially over all WM representations in turn, in order to maintain their activation and counteract time-based decay. Here again, as with SOB-CS, the reallocation of attention is, to a certain extent, already presumed; however, in this model, the emphasis is on the rapid, serial execution of that reallocation process and is termed "refreshing."

Our own viewpoint in this paper derives historically from, although is not strongly attached to, the multiple-entry, modular (MEM) framework (Johnson, 1983; Johnson, 1992; Johnson & Hirst, 1993). This includes our default conceptualization and operationalization of a refreshing process (i.e., the more intentional form of refreshing) as described in MEM and previously studied empirically in MEM-associated studies. In MEM, refreshing, as well as the other subprocesses described as part of MEM's reflective process subsystem (e.g., noting, reactivating, shifting), are already assumed to recruit attention. As a process-based framework, MEM does not explicitly include an "attention" element (or one for memory); rather, each putative subprocess in the framework is assumed to involve the allocation of attention in some way (and to create its own memory records), and subprocesses differ according to the manner and degree in which they allocate those resources. As such, the MEM version of refreshing is already highly compatible with the reallocation account we proposed above. We should note that MEM differs from SOB-CS and TBRS in that it makes no assumptions as to an underlying decay or interference mechanism; in fact, MEM generally is less of a formal model than those and more of a conceptual framework – a set of tools for articulating questions and forming hypotheses – and as such is less inclined towards making strong arguments for particular mechanistic implementations.

The debate between SOB-CS and TBRS has ranged over several years and multiple publications, with a full review of their implications perhaps beyond the scope of the present empirical investigation; and we would not be so bold as to claim we could resolve the debate between those models (or any of the various other WM models that exist) with just this single study. However, based on our finding of an apparent equivalence between refreshing and removal (and our suggestion that they both be reframed as a reallocation of mental attention), we might offer the interpretation that the question of refreshing vs. removal is in fact a distraction for purposes of comparing such models. Based on our current understanding, and in keeping with the spirit of the MEM framework that seeks largely to delineate a useful lexicon of mental processes to help further additional theoretical and empirical investigations, we see no major barrier to reframing SOB-CS, TBRS, or most other models in terms of attentional reallocation instead of refreshing/removal. Doing so would primarily result in shifting the burden of proof back to more fundamental underlying assumptions about basic mental mechanisms of time-based decay versus interference. Resolving differences in those low-level assumptions is admittedly a thornier problem than the more operationalizable process level of refreshing versus removal. We do not believe the present results make a firm statement in favor of either decay or interference accounts; we interpret them mainly to suggest that the question of "refreshing" versus "removal" may not be the ideal battleground on which to wage that war, and thus, for the time being, we must leave further resolution of the decay-versus-interference debate to the ingenuity of future investigators.

Lastly, the reallocation account also makes it somewhat easier to reconcile some of the awkward or counterintuitive aspects of framing WM phenomena starkly in terms of either refreshing or removal. For example, in a single-item directed-forgetting procedure, what is refreshed, or what happens when all contents of WM are removed? In the reallocation account, it is easy enough to contend that those attentional resources are simply strategically redirected to other representations within the participant's mental field of view, which may include representations outside the confines of the WM task proper – their upcoming exam, wondering how much longer this experiment will last, and so on. A final appeal of this reallocation account is that it is potentially easier to integrate with other conceptualizations of central attention and executive function, including visual attention processes that draw upon these central resources and interact/interfere with WM representations. Again, a complete exploration of this reallocation account would be better suited to a full review paper, but for now, we simply wish to raise the possibility that it represents a reasonable compromise between the refresh-centric or removal-centric accounts, and usage of the reallocation vocabulary may be helpful for clarifying and focusing the discourse surrounding differing theoretical views on WM.

4.3. Summary

In the current investigation, we sought to identify a behavioral distinction between the putative WM processes of refreshing and removal, if indeed such a distinction exists. Our results did not support any such distinction in a carefully controlled comparison, suggesting that there is no need to maintain that theoretical separation. While these separate terms may still have descriptive value in the context of referring to certain classes of cognitive tasks or patterns of experimental outcomes, we contend that it may be more fruitful to frame both concepts in terms of an overarching process of reallocating mental or reflective attention, and that further exploration of this reallocation-based perspective could be a promising avenue for future investigations.

Author note

This work was supported in part by NSF/EPSCoR grant 1632849 and by the National Institute of General Medical Sciences of the National Institutes of Health [grant number P20 GM130461] and the Rural Drug Addiction Research Center at the University of Nebraska-Lincoln. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the University of Nebraska.

We thank Arianna Dye for assistance with data collection and analysis.

Supplementary data: All data and code used in this study can be downloaded from: https://osf.io/azy8d/ (DOI: 10.17605/OSF.IO/AZY8D).

Declaration of Competing Interest

None.

References

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445–459. https://doi.org/10.3758/BF03193014.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83.

Barrouillet, P., & Camos, V. (2007). The time-based resource-sharing model of working memory. In N. Osaka, R. H. Logie, & M. D'Esposito (Eds.), *The cognitive neuroscience of working memory* (pp. 59–80). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198570394.003.0004.

Barrouillet, P., De Paepe, A., & Langerock, N. (2012). Time causes forgetting from working memory. *Psychonomic Bulletin & Review, 19*(1), 87–92. https://doi.org/10.3758/s13423-011-0192-8.

Camos, V., Johnson, M. R., Loaiza, V., Portrat, S., Souza, A., & Vergauwe, E. (2018). What is attentional refreshing in working memory? *Annals of the New York Academy of Sciences, 1424*(1), 19–32. https://doi.org/10.1111/nyas.13616.

Chun, M. M., & Johnson, M. K. (2011). Memory: Enduring traces of perceptual and reflective attention. *Neuron, 72*, 520–535. https://doi.org/10.1016/j.neuron.2011.10.026.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163–191. https://doi.org/10.1037/0033-2909.104.2.163.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57. https://doi.org/10.1177/0963721409359277.

Cowan, N., Elliot, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*, 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001.

Ecker, U. K. H., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory and Language, 74*, 1–15. https://doi.org/10.1016/j.jml.2014.03.006.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9*(1), 55–79. https://doi.org/10.3758/BF03196257.

JASP Team. (2018). *JASP (Version 0.9) [computer software]*.

Johnson, M. K. (1983). A multiple-entry, modular memory system. In G. H. Bower (Ed.), *Vol. 17. The psychology of learning and motivation: Advances in research and theory* (pp. 81–123). New York: Academic Press.

Johnson, M. K. (1992). MEM: Mechanisms of recollection. *Journal of Cognitive Neuroscience, 4*, 268–280. https://doi.org/10.1162/jocn.1992.4.2.258.

Johnson, M. K., & Hirst, W. (1993). MEM: Memory subsystems as processes. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories of memory* (pp. 241–286). East Sussex, England: Erlbaum.

Johnson, M. K., Reeder, J. A., Raye, C. L., & Mitchell, K. J. (2002). Second thoughts versus second looks: An age-related deficit in reflectively refreshing just-activated information. *Psychological Science, 13*(1), 64–67. https://doi.org/10.1111/1467-9280.00411.

Johnson, M. R., & Johnson, M. K. (2009). Top-down enhancement and suppression of activity in category-selective extrastriate cortex from an act of reflective attention. *Journal of Cognitive Neuroscience, 21*(12), 2320–2327. https://doi.org/10.1162/jocn.2008.21183.

Johnson, M. R., McCarthy, G., Muller, K. A., Brudner, S. N., & Johnson, M. K. (2015). Electrophysiological correlates of refreshing: Event-related potentials associated with directing reflective attention to face, scene, or word representations. *Journal of Cognitive Neuroscience, 27*, 1823–1839. https://doi.org/10.1162/jocn_a_00823.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General, 130*(2), 169–183. https://doi.org/10.1037/0096-3445.130.2.169.

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding attended information in short-term memory: An EEG study. *Journal of Cognitive Neuroscience, 25*(1), 127–142. https://doi.org/10.1162/jocn_a_00305.

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the nonexistent problem of decay. *Psychological Review, 122*(4), 674–699. https://doi.org/10.1037/a0039684.

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences, 13*(3), 120–126. https://doi.org/10.1016/j.tics.2008.12.003.

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience, 24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140.

Lewis-Peacock, J. A., Kessler, Y., & Oberauer, K. (2018). The removal of information from working memory. *Annals of the New York Academy of Sciences, 1424*(1), 33–44. https://doi.org/10.1111/nyas.13714.

Lintz, E. N., Lim, P. C., & Johnson, M. R. (2020). *A new tool for equating lexical stimuli across experimental conditions*. Manuscript submitted for publication. Department of Psychology, University of Nebraska-Lincoln. https://doi.org/10.31234/osf.io/64yfw.

MacLeod, C. M. (1998). Directed forgetting. In J. M. Golding, & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Morey, C. C., & Cowan, N. (2018). Can we distinguish three maintenance processes in working memory? *Annals of the New York Academy of Sciences, 1424*(1), 45–51. https://doi.org/10.1111/nyas.13925.

Muther, W. S. (1965). Erasure or partitioning in short-term memory. *Psychonomic Science, 3*(1−12), 429–430. https://doi.org/10.3758/BF03343215.

Oberauer, K. (2001). Removing irrelevant information from working memory: A cognitive aging study with the modified Sternberg task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4), 948–957. https://doi.org/10.1037/0278-7393.27.4.948.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review, 19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4.

Pierce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017.

Portrat, S., Barrouillet, P., & Camos, V. (2008). Time-related decay or interference-based forgetting in working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1561–1564. https://doi.org/10.1037/a0013356.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology, 55A*, 1339–1362. https://doi.org/10.1080/02724980244000099.

Raye, C. L., Johnson, M. K., Mitchell, K. J., Greene, E. J., & Johnson, M. R. (2007). Refreshing: A minimal executive function. *Cortex, 43*, 135–145. https://doi.org/10.1016/S0010-9452(08)70451-9.

Raye, C. L., Johnson, M. K., Mitchell, K. J., Reeder, J. A., & Greene, E. J. (2002). Neuroimaging a single thought: Dorsolateral PFC activity associated with refreshing just-activated information. *NeuroImage, 15*, 447–453. https://doi.org/10.1006/nimg.2001.0983.

Shepherdson, P., & Oberauer, K. (2018). Pruning representations in a distributed model of working memory: A mechanism for refreshing and removal. *Annals of the New York Academy of Sciences, 1424*, 221–238. https://doi.org/10.1111/nyas.13659.

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics, 78*, 1839–1860. https://doi.org/10.3758/s13414-016-1108-5.

Souza, A. S., & Vergauwe, E. (2018). Unravelling the intersections between consolidation, refreshing, and removal. *Annals of the New York Academy of Sciences, 1424*, 5–7. https://doi.org/10.1111/nyas.13943.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7.

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53*(1), 5–13. https://doi.org/10.1037/0022-3514.53.1.5.

Williams, M., & Woodman, G. F. (2012). Directed forgetting and directed remembering in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1206–1220. https://doi.org/10.1037/a0027389.