Offloading Memory Leaves Us Vulnerable to Memory Manipulation

E. F. Risko*, M. O. Kelly, P. Patel, & C. Gaspar

University of Waterloo

*Correspondence to: E. F. R. at efrisko@uwaterloo.ca

Data and analysis code are available here: https://osf.io/3ce6u/

Funding: This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), an Early Researcher Award from the Province of Ontario and funding from the Canada Foundation for Innovation, Ontario Research Fund and Canada Research Chairs program to E.F.R.

This is a post-peer-review, pre-copyedit version of an article published in Cognition. The final authenticated version is available online at: https://www.sciencedirect.com/science/article/pii/S0010027719301076?via%3Dihub

Abstract

We often offload memory demands onto external artefacts (e.g., smartphones). While this practice allows us to subvert the limitations of our biological memory, storing memories externally exposes them to manipulation. To examine the impact of such manipulation, we report three experiments, two of which were pre-registered. Individuals performed a memory task where they could offload to-be-recalled information to an external store and on a critical trial, we surreptitiously manipulated the information in that store. Results demonstrated that individuals rarely noticed this manipulation. In addition, when individuals had information inserted into their external memory stores, they often encoded it into their biological memory, thereby leading to the creation of a false memory. The reported results highlight one of the cognitive consequences of offloading our memory to external artefacts.

Offloading Memory Leaves Us Vulnerable to Memory Manipulation

Humans have long sought to offload demands on their memory by using external artefacts (Clark, 2010a; Donald, 1991; Nestojko, Finley & Roediger, 2013; Risko & Gilbert, 2016). Nevertheless, we lack a deep understanding of the principles governing this distributed form of remembering. The recent proliferation and increasing availability of mass storage devices presents our species with a remarkable opportunity to store large amounts of easily accessible information that is immune from the vicissitudes of our biological memory; however, offloading memory creates its own set of risks (Carr, 2008; Eskritt & Ma, 2014; Risko & Gilbert, 2016; Sparrow, Liu & Wegner, 2011). One such risk is that offloading memory to an external location exposes it to manipulation by a third party (Clark, 2010b; Sterelny, 2004). For example, an agent could surreptitiously alter our "memory" by manipulating the information in our external memory stores. Understanding how individuals respond to this kind of manipulation would provide insight into how individuals manage the unique challenges presented by distributing memory demands over internal and external spaces (e.g., transactive memory systems; Arango-Munoz, 2013; Ferguson, McLean & Risko, 2015; Gilbert, 2015; Risko, Ferguson & McLean, 2016; Risko & Dunn, 2015; Sparrow et al., 2011; Ward, 2013; Wegner, 1995; Sutton, Harris, Keil & Barnier, 2010; Storm, Stone & Benjamin, 2017). Provided the ubiquity of offloading as a means of remembering, understanding our susceptibility to external memory manipulation, the factors that modulate it, and the impact of such manipulation on our biological memory is needed. To this end, we report three experiments examining external memory manipulation using a novel paradigm.

In the context of a distributed cognitive system, individuals face various challenges (Arango-Munoz, 2013). This includes deciding whether to solve a cognitive problem relying on internal resources or external resources (or a mix of the two; i.e., the "selection" problem; Arango-

Munoz, 2013). Much recent work has focused on this aspect of cognitive offloading (i.e., how do we decide to offload to-be-remembered information rather than store it internally; Cherkaoui, & Gilbert, 2017; Gilbert, 2015a; 2015b; Risko, Medimorec, Chisholm & Kingstone, 2014; Risko & Dunn, 2015; Risko & Gilbert, 2016). In addition to deciding whether to rely on an external resource, individuals also have to decide whether to rely on or "endorse" the information provided by that external resource (i.e., the "endorsement" problem; Arango-Munoz, 2013). For example, in a memory context, if an individual stores some to-be-remembered information in an external location, at the time of retrieval from that external store the individual needs to endorse that information as the original to-be-remembered information. Previous research suggests that individuals may have great difficulty detecting manipulations of their external memory stores or, in other words, solving this endorsement problem.

Sparrow et al. (2011) demonstrated that offloading to-be-remembered information (i.e., storing information externally with the expectation of future access to that store) leads to a compromised ability to recall that information unaided (see also Eskritt & Ma, 2014; Kelly & Risko, in press). If information is poorly encoded because we expect to have access to it via an external store, then our ability to distinguish between legitimate and illegitimate information in that external memory store may be compromised (e.g., impaired discrepancy detection; Tousignant, Hall & Loftus, 1986). A second source of difficulty would likely emerge from the metacognitions we have about both our biological and external memory stores (e.g., Mazzoni & Kirsch, 2002). If individuals already believe that offloaded information is associated with compromised biological memory, then this would provide a plausible rationale for the absence of experience of recall during retrieval from the external memory store (i.e., a "forgetting rationale"; Scorobia, Lynn, Hessen & Fisico, 2007). This rationale could be further bolstered by our belief in

the reliability of common external memory stores (e.g., computer files; Risko & Dunn, 2015; Risko & Gilbert, 2016; Storm & Stone, 2015; Storm et al., 2017). Together, the lack of memorial support for detecting discrepancies, a plausible reason that lack of memory is not diagnostic of manipulation, and an extant belief in the reliability of external memory stores could all conspire to render us deeply susceptible to their manipulation.

Present Investigation

In the reported experiments, individuals were presented with a list of to-be-remembered words, one at a time, and were told to type this information into an external memory store, a file on a computer. After the list was presented, the individual performed a distractor task with their file out of view. Following the distractor task, the individual was given back access to their external memory store and was asked to recall the words that were presented previously. This paradigm mimics our typical experience offloading memory. We encounter information that we would like to recall in the future, we commit it to an external memory store, and when needed, we retrieve the information from that external location. In our experiments, individuals perform a number of such trials, thus, developing trust in the external store. The critical manipulation occurs during the final trial. While the individual is performing the distractor task, we altered the computer file that was functioning as the individual's external memory.

Experiment 1

In Experiment 1 (pre-registered at https://osf.io/v8bzr/), individuals performed the task described above. The manipulation on the final trial consisted of inserting a new word into their list. This inserted word was not semantically, phonologically, orthographically or otherwise meaningfully related to the other items in the list (Roediger & McDermott, 1995; Watson, Balota & Roediger, 2003). This, along with the effective encoding task (i.e., typing words into a file;

Forrin, MacLeod, & Ozubko, 2012) and the short time between encoding and "retrieval" (e.g., less than a minute) should all enhance an individual's chances of detecting the insertion, thus allowing us to provide a strong test of the depth of our susceptibility. We were also interested in beginning to identify the factors relating to our ability to notice manipulations of our external memory stores. To this end, we manipulated where we inserted the word within the individual's stored list. Individuals typically show a greater likelihood of successful retrieval for items presented at the beginning of a list, relative to items in the middle (Glanzer & Cunitz, 1966; Suprenant & Neath, 2000). In addition, individuals are thought to pay greater attention to the initial items in a list, relative to later items (e.g., Brown, Preece, & Hulme, 2000; Farrell & Lewandowsky, 2002). If the likelihood of successfully retrieving the original information and/or our attentiveness modulates our susceptibility to the alteration of our external memories, then manipulations located at the beginning of an offloaded list might be more likely to be noticed than manipulations located in the middle of an offloaded list.

We indexed the extent to which individuals noticed the manipulation of their external memories using a number of different measures. This included whether participants, during the recall phase, reported the words as they were originally presented or as they were represented in the individual's external memory store (i.e., with the inserted word). During the recall phase, participants were also asked to rate their confidence (from 0% to 100%) that each of their reported words had, in fact, been presented. Following the recall phase, we asked individuals three increasingly specific open-ended questions about the lists of words that they had just been asked to remember. We classified their responses based on the extent to which they expressed knowledge of the manipulation of their external memory. Finally, we explicitly disclosed to participants that

we may have added a word to their final list and asked them to select the one that may have been added.

Experiment 1

Method

Participants. Seventy-five undergraduate psychology students from the University of Waterloo participated for course credit. We had to remove three participants, all of whom did not properly complete the experiment. We planned to stop data collection once we had 72 useable participants divided equally across the two between subject conditions (36 in each condition). This was based on having .80 power to detect a 40% difference in participants' ability to pick out the inserted word on the final question of the post-task questionnaire. Using the z test "Proportions: Difference between two independent proportions" in G*Power 3.1.9.2 (alpha = .05), this would require at least 48 (24 in each condition) participants (Faul, Erdfelder, Lang, & Buchner, 2007).

Apparatus. The participant and the researcher sat at their own workstations, separated by a partition. The researcher workstation consisted of two computers with two corresponding monitors. The participant's workstation included a keyboard and mouse for one monitor (the workspace monitor) and another monitor (the display monitor) which displayed the instructions throughout the session. Word lists were presented originally through speakers at the participant workstation. The workstation also included pens and paper sheets for the recall tasks and post-task questionnaire. The monitors at the participant workstation were connected to the computers and monitors at the researcher workstation, which allowed the researcher to remotely control the monitors at the participant workstation.

Stimuli. We used four word lists. The words within each of the lists (and between) were not meaningfully related to one another and were drawn from the SenticNet 4 word corpus

(Cambria, Poria, Bajpai & Schuller, 2016). Audio of each word was individually recorded. In order to avoid a change in list length being a salient cue to the insertion of the words, the lists varied in length (i.e., 18, 20, 20, and 22). We designated one word for each list as the inserted word that the researcher would insert when that list appeared on the final trial. When these words were inserted into the middle position, this corresponded to the 9th, 10th and 11th position for the 18, 20, and 22 word lists respectively. The mean log word frequency of the non-inserted/actually presented words was 7.22 (using frequency count from S*UBTLEX-UK*; Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) and the mean word length was 6.24 letters. The inserted words (which were the same across participants) had a mean log word frequency of 8.97 and a mean word length was 7.50 letters. These word types did not differ in mean word length, $t(3.57) = 1.87$, $p = .143$, $d = 0.75$, but did in word frequency, $t(4.28) = 2.97$, $p = .038$, $d = .87$.[1] When the lists were used on non-manipulated trials (1 to 3), the "inserted" words were not inserted and thus, list lengths were 17, 19, 19, and 21. These "inserted" words are referred to as "yoked words" below (i.e., words that would have been inserted, had their corresponding list been the final trial list). The lists were counterbalanced across participants, such that each list was presented on each trial equally often, however, the word order within the lists was held constant across participants. This facilitated the timely insertion of the word on the fourth trial (i.e., the researcher inserted the word after a particular word rather than having to locate the insertion point by counting).

Procedure. The participant was seated approximately 50 centimeters in front of the two side-by-side monitors (workspace and display monitors). Following consent, the participant followed instructions given by the display monitor and the researcher throughout the experiment. Each trial began with an encoding phase wherein the list of to-be-remembered words was presented auditorily, one word at a time, separated by 4000 ms intervals. The researcher instructed the

participant to simultaneously type these presented words into the '.txt' file on the workspace monitor. The researcher remotely turned off the participant's workspace monitor once all of the words from the list were presented. While the workspace monitor was turned off, the participant performed a distractor task that consisted of counting backwards by sevens from a random four-digit number aloud for 30 seconds. The researcher monitored accuracy to ensure that the participant was engaged, but they did not record it. After the distractor task, the researcher remotely turned back on the workspace monitor to give the participant access to their file. The participant was asked to recall the words that they had heard onto a piece of paper. In addition, participants were asked to provide a rating beside each word with respect to how confident they were that the word was presented originally on a scale of 0% to 100% confidence. During this recall phase, the participant was allowed to refer to the '.txt' file. There were four trials in total. On the fourth and final trial, when the participant was performing the distractor task, the researcher inserted a word into their typed list. The word was inserted into either the first or middle position of the list.

After the final trial's recall phase, participants completed a questionnaire that consisted of four questions to determine the extent to which they noticed the insertion of the word. The questions were administered sequentially on the participant's workstation monitor. Each question increased in specificity regarding the manipulation. Question 1 asked, "Did anything stand out to you about the experiment? If so, then what?" Question 2 asked, "Did anything stand out in any of the lists? If so, then what?" Question 3 asked, "Did anything stand out in the final list? If so, then what?" Finally, a fourth and final question asked "On the last list we may have added a word that was not presented originally. Please review the list and type the word that you think was added." Participants were given their final recalled list to refer to for this question. Thereafter, the

researcher fully debriefed the participant regarding the purpose of the study and the deception used.

Results

Data and analysis code for Experiments 1 to 3 are available here: https://osf.io/3ce6u/. Two naïve coders individually coded the answers to the post-task questionnaire as 0, 0.5, or 1, indicating "no knowledge", "partial knowledge", or "full knowledge" for each of the first three questions. Initial agreement was 90%. Disagreements between coders were reconciled by having them reach agreement on the classification without researcher involvement. Coders did not need to code answers to the fourth question, as they were not subjective in nature. There were ten (0.18% of total) cases wherein participants did not enter a word into their external memory store, 33 (0.60% of total) cases wherein participants did not report a word during recall, and three (0.05% of total) cases wherein participants did not provide confidence ratings for reported words during recall. As such, these cases could not be included in the confidence analyses. Finally, all confidence intervals reported throughout are bias corrected accelerated bootstrap 95% confidence intervals using 10000 replications. All paired non-parametric tests are corrected for potential ties using "wilcox.exact" from the *exactRankTests* R package (Hothom & Hornik, 2019).

Recall with External Memory Store (Trial 4). On the manipulated list (fourth trial), 100% of participants wrote down the inserted word. No participant inserted the yoked word (i.e., words that would have been inserted had the corresponding list appeared as the final trial list; see *Stimuli* above) on a non-manipulated list (i.e., trials 1-3). As such, the pre-registered analysis comparing inserted items to yoked items was not conducted.

Participants expressed high confidence that items inserted into their external memory had been presented originally. In the middle position condition, mean confidence for inserted words was 79% [CI: 68%, 87%] and in the first position condition, mean confidence in the inserted word

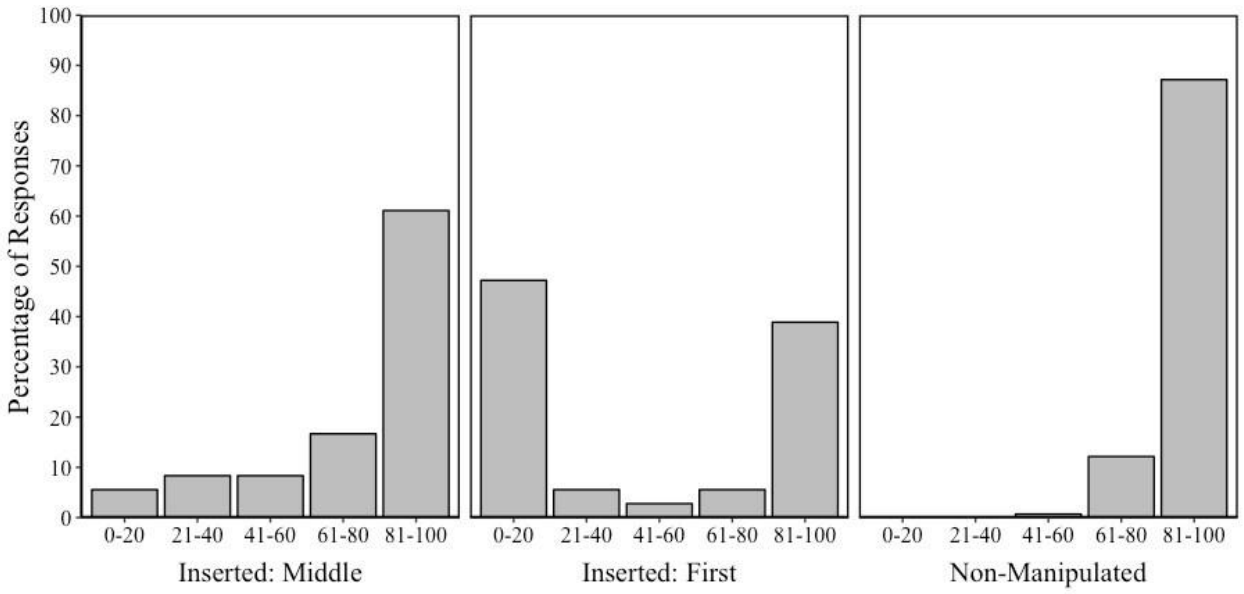

Figure 1. *The percentage of confidence responses that fell into each of five consecutive 20% bins in the recall with external memory store task as a function of word type and condition in Experiment 1. The non-manipulated panel contains the data from both middle and first position conditions.*

was 46% [CI: 31%, 61%]. This difference was significant with both a parametric, $t(58.13) = 3.67$, $p < .001$, $d = 0.87$, and non-parametric test, $W = 415$, $p = .006$. As illustrated in Figure 1, the difference in confidence across conditions appears largely to be a function of the majority of participants in the middle position condition being highly confident (> 80%) in the inserted word, whereas in the first position condition, a similar number of participants expressed high confidence (> 80%) as low confidence (≤ 20%). If we compare confidence in each condition to 50%, then those in the first position condition did not differ significantly, $t(35) = 0.54$, $p = .593$, whereas

those in the middle position condition did, $t(35) = 6.14$, $p < .001$ (these analyses were not pre-registered). In addition, if we consider 50% to be the threshold between "yes" and "no" responses, then 47% in the first condition responded "yes" and 86% in the middle condition responded "yes" to the inserted word.

Overall confidence for words that were actually presented on trial 4 was 89% [CI: 87%, 91%]. Participants with words inserted into the first position of their typed lists were significantly less confident that the word had been presented than they were that the other final list (non-manipulated) words had been presented, $t(35) = 5.91$, $p < .001$, $d = 0.98$, that the non-manipulated words across all lists had been presented (all non-manipulated: 89% [CI: 86%, 91%]), $t(35) = 5.74$, $p < .001$, $d = 0.97$, and that the item adjacent (one after) the inserted item had been presented (adjacent: 82% [CI: 71%, 89%]), $t(35) = 3.84$, $p < .001$, $d = 0.64$ (the latter analysis was not pre-registered). A qualitatively similar pattern emerged when using non-parametric tests (these analyses were not pre-registered). Lastly, participants had significantly lower confidence in the inserted item when it appeared in the first position than non-manipulated items when they appeared in the first position (non-manipulated first: 70% [CI: 59%, 79%]), $t(60.89) = 2.61$, $p = .011$, $d = 0.62$ and with a non-parametric test, $W = 443.50$, $p = .028$ (these analyses were not pre-registered). It is important to note here that the confidence in the first item when it was not inserted (i.e., the first item on the fourth trial in the middle position condition) was lower than other non-manipulated items, thus the lower confidence in the inserted item when it appeared in the first position is likely partially due to a generally lower confidence in that item.

Participants with words inserted into the middle position of their typed list were less confident that the word had been presented originally than they were that other final list (non-manipulated) words had been presented, $t(35) = 2.54$, $p = .016$, $d = 0.42$, though this difference

was not significant when using a non-parametric test, $V = 215$, $p = .163$ (the latter analysis was not pre-registered). We found a similar pattern when we compared confidence in the inserted word to confidence in non-manipulated words on all of the lists. That is, there was a significant difference using a parametric test (all non-manipulated: 90% [CI: 87%, 92%]), $t(35) = 2.40$, $p = .022$, $d = 0.40$, that was not significant with a non-parametric test, $V = 282$, $p = .432$ (the latter analysis was not pre-registered). Participants had significantly higher confidence in the item adjacent (one after) the inserted item compared to the inserted item (adjacent: 96% [CI: 91%, 98%]), $t(35) = 3.22$, $p = .003$, $d = .54$. A qualitatively similar pattern emerged when using non-parametric tests (these analyses were not pre-registered). Lastly, confidence in the inserted item when it appeared in the middle position was not significantly different than confidence in non-manipulated items when they appeared in the middle position (non-manipulated middle: 81% [CI: 68%, 90%]), $t(69.09) = 0.31$, $p = .757$, $d = 0.07$, and with a non-parametric test, $W = 751.50$, $p = .209$ (these analyses were not pre-registered).

It is important to note that a number of participants (17 in the first position condition; two in the middle position condition) recalled the inserted word during the recall phase but with 0% confidence that it had been presented. This suggests that the interpretation of the participant's act of recall (i.e., writing the word down on the recall sheet) in and of itself should proceed cautiously (i.e., why write the word down indicating that it had been presented to only indicate 0% confidence that it had been presented?). It is possible that participants relied on their external memory store for what to report during the recall phase, but when reporting confidence, relied on their subjective experience of the item (e.g., "I don't recall this word at all but because it's in there, it must have been presented"). Alternatively, participants might have thought they were "supposed" to write whatever was in their external memory store during recall but on other measures expressed their

"true" beliefs. Lastly, it is important to note that there was little to no cost to writing down all the words and this might have encouraged a more liberal strategy when deciding to write the word during recall. Whichever is the case, it underlines the importance of using multiple measures of "noticing" the manipulation as we have done here.

Post-Task Questionnaire. Participants showed little to no explicit knowledge of the manipulation of their external memory during questioning particularly in the middle position condition. Comparisons across conditions are based on coders' numeric ratings (0: no knowledge; 0.5: partial knowledge; 1: full knowledge). For descriptive purposes we present the percentage of responses in each category as a function of condition in Table 1 for the three open-ended questions. Participants with words inserted into the first position of their typed list expressed significantly more knowledge of the manipulation than participants with words inserted into the middle position on question 1 (first: .19 [CI: .06, .31] vs. middle: .04 [CI: 0, .08]), $t(43.41) = 2.16$, $p = .037$, $d = 0.51$, though this was not significant when using a non-parametric test, $W = 731$, $p = .087$. There was no difference as a function of condition on question 2 (first: .14 [CI: .03, .25] vs. middle: .11 [CI: .03, .22]) with either a parametric, $t(68.04) = .36$, $p = .717$, $d = 0.09$, or non-parametric test, $W = 653$, $p = .853$. There was a significant difference in the amount of knowledge expressed on the third question, $t(58.53) = 3.13$, $p = .003$, $d = 0.74$, such that participants with words inserted into the first position of their typed list expressed significantly more knowledge of the manipulation than participants with words inserted in the middle position (first: .40 [CI: .25, .54] vs. middle: .11 [CI: .03, .22]). This difference was also significant when using a non-parametric test, $W = 855$, $p = .004$. On question 4, participants with words inserted into the first position picked out the word 53% [CI: 33%, 67%] of the time compared to 28% [CI: 14%, 42%] of the time for participants with words inserted into the middle position. This difference was significant,

$\chi^2(1) = 4.68$, $p = .031$. Participants were significantly above chance (5%) in picking out the inserted word in both the middle position, $t(35) = 3.01$, $p = .005$, $d = 0.51$, and first position conditions, $t(35) = 5.66$, $p < .001$, $d = 0.96$. Participants clearly had difficulty picking out the word we had inserted into their external memory stores. It is important to note that an error (i.e., not picking the inserted word) on this task indicates that participants picked out a word that had actually been presented.

Table 1. The percentage of responses that fell into each classification across experiments and conditions on the three open-ended questions of the post-task questionnaire. In Experiment 1 and 2 responses were coded as expressing "No Knowledge," "Partial Knowledge," or "Full Knowledge." In Experiment 3 responses were coded as "No Knowledge" and "Some Knowledge."

	Question 1			Question 2			Question 3		
	No	Partial/ Some	Full	No	Partial/ Some	Full	No	Partial/ Some	Full
Experiment 1									
First	81	0	19	86	0	14	56	8	36
Middle	92	8	0	86	6	8	86	6	8
Experiment 2									
Middle	85	7.5	7.5	87.5	7.5	5	80	17.5	2.5
Experiment 3									
First	87	13	NA	95	5	NA	95	5	NA
Middle	97	3	NA	100	0	NA	95	5	NA

Combined measures of lack of awareness. We computed two aggregate measures of lack of awareness by combining our various measures of "noticing" the manipulation of the external memory store. For the first aggregate measure, participants were defined as "unaware" of the manipulation if they reported the inserted word during recall, indicated 100% confidence in the inserted word having been presented originally, and indicated no knowledge on the initial three questions of the post-task questionnaire. These criteria were met by 25% of participants with words inserted into the first position and 42% of participants with words inserted into the middle position.

This difference was not significant, $\chi^2(1) = 2.25$, $p = .134$.The second aggregate measure used all three criteria above, adding that the participant also failed to pick out the inserted word when told that a word may have been added to their final list (i.e., the final question of the post-task questionnaire). These criteria were met by 19% of participants with words inserted into the first position and 33% of participants with words inserted into the middle position. This difference was not significant, $\chi^2(1) = 1.79$, $p = .181$. Thus, between approximately one-quarter to one-half of participants showed a complete lack of awareness across multiple measures. There was no effect of location on either of these metrics of the manipulation.

Discussion

Experiment 1 provided a number of insights into our susceptibility to the manipulation of our external memory stores. When asked to write down the words that had been presented, all participants wrote down the inserted word on the critical fourth trial. Participants also reported relatively high confidence in the inserted words, were limited in their ability to explicitly express knowledge of the manipulation of their external memory (i.e., the majority on all questions expressed "no" knowledge), and struggled to pick out that word when pressed (though, they did so above chance levels). Lastly, on a number of measures (not all) the location of the manipulation impacted participant's responses. For example, participants were less confident that the inserted item had been presented, expressed more explicit knowledge of that insertion, and were better able to pick out the inserted item when it was inserted in the first position in the list. At least some of this location effect seemed to be due to a general lack of confidence in the first items of the lists. Lastly, and importantly, the results also clearly demonstrate, in terms of confidence, that a participant's experience of the inserted item differed from words that were actually presented (i.e., it was lower), though this was not always the case (e.g., the middle position items).

Experiment 2

In Experiment 2, we examined whether the information that we surreptitiously inserted into an individual's external memory store would become part of their biological memories for the event. Previous research on the memorial consequences of misleading post-event information (e.g., Loftus, 2005) and false evidence (e.g., Nash & Wade, 2009; Nash, Wade, & Lindsay, 2009; Strange, Gerrie, & Garry, 2005) suggests that the manipulation of an individual's external memory could lead to the formation of a false biological memory at a high rate. For example, plausibility is thought to play an important role in false memory formation (e.g., Pezdek, Finger & Hodge, 1997) and, as noted above, an item inserted into an external memory store is likely to appear highly plausible by its mere presence in that store. To test this idea, Experiment 2 added a recognition test following the recall phase of the final list and an additional 30 second distractor task. During this test, participants were presented with words that had been presented during the experiment (i.e., the final/fourth list and the third and first lists), words that had not been presented (i.e., new words) and, critically, the word that had been inserted into the final list. Participants were asked to report whether a given word had appeared, specifically, on the final list. Importantly, participants did not have access to their external memory stores during this surprise recognition test and had to rely completely on their biological memory.

Methods

Participants. Forty undergraduate students from the University of Waterloo participated for either course credit or ten Canadian dollars. We planned to stop data collection once we had 40 useable participants. This was based on the sample sizes used in Roediger and McDermott

(1995) of 36 and 30, plus the added constraint of completing the counterbalance of word lists across conditions.

The *Apparatus, Recall with External Memory Store, and Post-task Questionnaire,* used were the same as in Experiment 1.

Stimuli. The stimuli were the same as in Experiment 1, however, we added another list so that one of the five lists would not be presented to participants and would instead act as the set of new words during the surprise final recognition task. This added list was 24 words long. The mean log word frequency of the non-inserted/actually presented words was 7.26 (using frequency count from S*UBTLEX-UK*; Van Heuven et al., 2014) and mean word length was 6.21 letters. The inserted words (which were the same across participants) had a mean log word frequency of 8.58 and a mean word length was 7.20 letters. These words did not differ in mean word length, $t(4.71) = 1.63$, $p = .168$, $d = 0.59$, but the difference in word frequency was marginal, $t(5.15) = 2.18$, $p = .080$, $d = 0.65$. The five lists were counterbalanced across participants, such that each list appeared on each trial (1 to 4) and acted as the unpresented list (used in the recognition with no external memory store) equally often.

Recognition with No External Memory Store. After the final trial, participants performed a recognition memory task on the display monitor. Words appeared randomly, one at a time. All words from the first, third, and fourth list positions were presented to participants on the recognition task, in addition to new words. New words included all words from the unpresented list and the "inserted" (i.e., yoked) words of list positions 1 to 3 (which as noted above were not presented). Crucially, the word that was inserted into the fourth and final list was also presented during the recognition task. Participants rated their confidence from 1 to 4 as: (1) Definitely not presented originally through the audio on the final list (i.e., list 4), (2) Probably not presented…

(3) Probably presented… (4) Definitely presented… (adapted from Roediger & McDermott, 1995).

Procedure. The procedure used in Experiment 2 was the same as in Experiment 1 with the following exceptions: (1) the final list manipulation was always the insertion of the word in the middle position, (2) after the final trial, participants performed another brief distractor task (counting backwards by sevens from a random four digit number for 30 seconds) and then (3) we administered the final recognition task, followed by the post-task questionnaire. The latter change is important to note (i.e., the post-task questionnaire now follows an intervening task) when making direct comparisons to the same task in Experiment 1.

Results

Participant answers to the post-task questionnaires were coded in the same manner as in Experiment 1 with initial agreement between coders of 90%. Disagreements were reconciled in the same manner as in Experiment 1. There were 16 (0.50% of total) cases wherein participants did not encode a word into their external memory storage, 24 (0.75% of total) cases wherein participants did not report a word during recall, and five (0.16% of total) cases wherein participants did not provide confidence ratings for reported words during recall. As such, these cases could not be included in confidence analyses.

Recall with External Memory Store (Trial 4), Post-Task Questionnaire and Combined Measures of Lack of Awareness. As in Experiment 1, on the manipulated list (trial 4), 100% of participants wrote down the inserted word during recall with their external store and no one inserted a yoked word on a non-manipulated list. Confidence in the inserted word in Experiment 2 (50% [CI: 38%, 62%]) was lower than the comparable (middle position) condition of Experiment 1 but similar to the first position condition of Experiment 1. Confidence did not differ from 50%

overall, $t(39) = 0$, $p = 1.00$, $d = 0$ and if we consider 50% to be the threshold between "yes" and "no" responses, then 53% responded "yes" to the inserted word. Again, participants were less confident that the inserted word had been presented originally than they were that other final list (non-manipulated) words had been presented (final list non-manipulated: 85% [CI: 81%, 88%]), $t(39) = 6.47$, $p < .001$, $d = 1.02$, and that all non-manipulated words had been presented (i.e., from trials 1 to 4; all non-manipulated: 85% [CI: 81%, 88%]), $t(39) = 6.25$, $p < .001$, $d = 0.99$. Participants had significantly higher confidence in the item adjacent (one after) the inserted item compared to the inserted item (adjacent: 78% [CI: 65%, 86%]), $t(39) = 4.56$, $p < .001$, $d = 0.72$. A qualitatively similar pattern emerged when using non-parametric tests.

On questions 1, 2, and 3 of the post-task questionnaire (see Table 1 for the percentage of responses in each category), participants expressed a limited about of knowledge that their external store had been manipulated. The mean knowledge was respectively .11 [CI: .04, .21], .09 [CI: .03, .18], and .11 [CI: .05, .19] for questions 1, 2, and 3. On question 4, 28% [CI: 13%, 40%] of participants were able to identify the manipulated word correctly. This was comparable to the middle position condition of Experiment 1 and differed significantly from chance (~ 4.88 %), $t(39) = 3.16$, $p = .003$, $d = 0.50$. Lastly, 25% of participants reported the inserted word during recall, indicated 100% confidence in the inserted word having been presented originally, and indicated no knowledge on the initial three questions of the post-task questionnaire, while 23% did the former and did not pick out the inserted word when told that a word may have been inserted into their list. These values were similar to Experiment 1.

Recognition without External Memory Store. Participant ratings were converted to yes/no answers (1 and 2 as "no"; 3 and 4 as "yes") and treated continuously (i.e., 1 to 4) in the analyses to follow (see Figure 2).

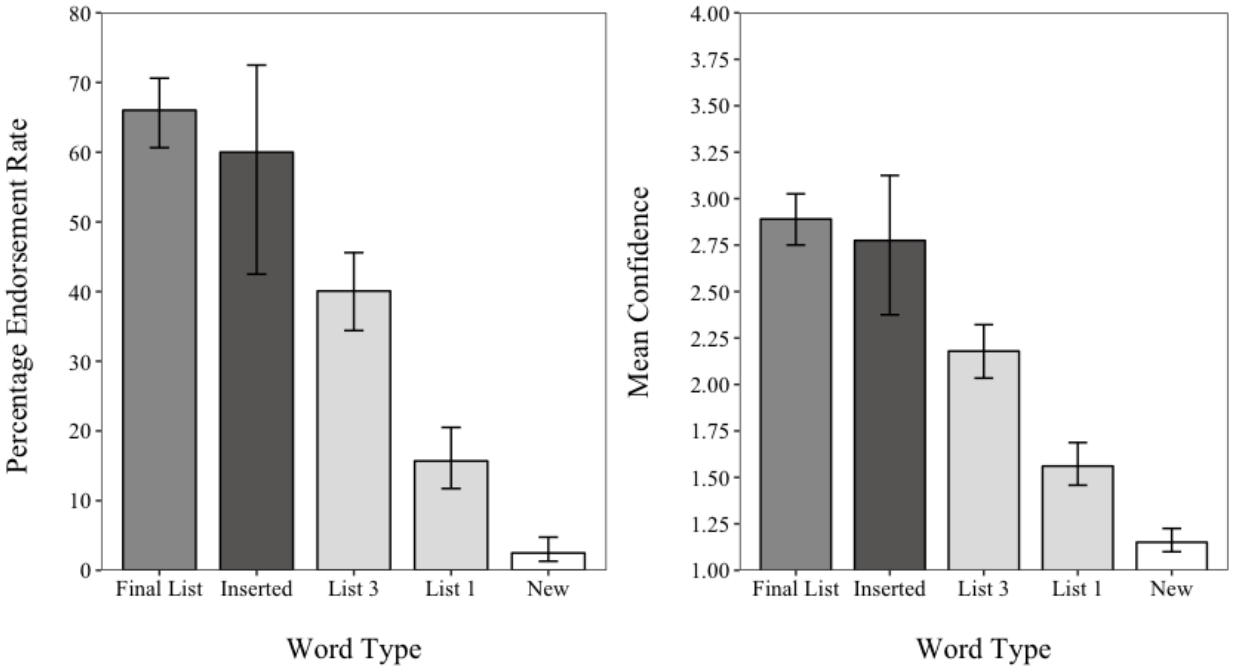

Figure 2. *Individual's responses in the final recognition test without the external memory store in Experiment 2 as a function of word type. Left panel: Percentage of "Yes" responses where "3" and "4" are coded as "Yes." Right panel: confidence responses treated as continuous (i.e., from 1 to 4). Error bars are bias corrected accelerated bootstrap confidence intervals using 10,000 replications.*

Participants exhibited clear evidence that the manipulation of their external memory store had altered their biological memory for the final list. Participants reported that the inserted word had been presented on the final list 60% [CI: 43%, 73%] of the time. The rate at which participants said "yes" to the inserted word was not statistically different from the rate at which participants said "yes" to words that were actually presented on the final list (66% [CI: 61%, 71%]), $t(39) = 0.77$, $p = .444$, $d = 0.12$. This was also true when participant responses were treated as continuous (inserted: 2.78 [CI: 2.38, 3.12] vs. list 4: 2.89 [CI: 2.75, 3.03]), $t(39) = 0.61$, $p = .547$, $d = 0.10$. The Bayes factor in support of the null in both cases provided modest support, $BF_{01(r=.707)} = 4.43$,

$BF_{01(r=.707)}$ = 4.93, for endorsement and confidence respectively. The distribution of responses across the four options was also similar for inserted words and words that had actually been presented (see Table 2). In both cases the majority of the responses fell into the fourth response option "Word was definitely presented…"

Table 2. Mean percentage of the total number of responses for each confidence rating (1 to 4) as a function of word type on the final recognition test without the external memory store of Experiment 2.

	Confidence Rating			
Word Type	1: Definitely Not	2: Probably Not	3: Probably	4: Definitely
Final List	23	11	20	46
Inserted	22.5	17.5	20	40
List 3	42	18	20	20
List 1	66	18	9	6
New	89	8	1	2

Note. For List 1 items, the total percentage by confidence rating sums to 99 due to rounding.

The percentage of "yes" responses to inserted words were also significantly higher than the percentage of "yes" responses to non-manipulated words in list 3, list 1 and new words (inserted: 60% [CI: 41%, 73%] vs. list 3: 40% [CI: 35%, 46%] vs. list 1: 16% [CI: 12%, 21%] vs. new: 2.5% [CI: 1%, 5%]), all *t*s ≥ 2.45, *p*s ≤ .019, *d*s ≥ 0.39. When these answers were treated as continuous (i.e., from 1 to 4), the results were similar (inserted: 2.78 [CI: 2.38, 3.10] vs. list 3: 2.18 [CI: 2.04, 2.32] vs. list 1: 1.56 [CI: 1.46, 1.69] vs. new: 1.15 [CI: 1.10, 1.22]), all *t*s ≥ 2.96, *p*s ≤ .005, *d*s ≥ 0.47. Lastly, pairwise differences between non-manipulated word types (i.e., list 4, list 3, list 1, and new) were all significant in the percentage of "yes" responses, all *t*s ≥ 5.77, *p*s ≤ .001, *d*s ≥ 0.91, and when answers were treated as continuous, all *t*s ≥ 7.36, *p*s ≤ .001, *d*s ≥ 1.16. A qualitatively similar pattern emerged when using non-parametric paired comparisons. A number of pairwise comparisons were reported in the above section. When applying a correction for multiple comparisons, the only pairwise comparisons that were no longer significant were the

comparisons between the inserted word and the list 3 words for endorsement $t(39) = 2.45$, $p = .190$, $d = 0.39$, and confidence considered continuously $t(39) = 2.96$, $p = .053$, $d = 0.47$.

Exploratory

Relation between Recall and Surprise Recognition. An important prediction of the discrepancy detection notion in the misinformation literature is that the likelihood of misinformation being integrated into memory should be greater if individuals fail to notice the "discrepancy" (Tousignant et al., 1986). In the present context, this predicts a relation between responses in the recall with external memory store (Trial 4) task and the recognition without external memory store test. As noted above, 100% of participants wrote down the inserted word during recall, thus we used confidence at recall as a predictor of responses in the surprise recognition task. We treated confidence during recall with external store (Trial 4) as binary (< 50% vs. ≥ 50%). Participants who reported confidence greater than or equal to 50% in the recall with external memory store (Trial 4) test were more likely to endorse the inserted item in the surprise recognition task (.76 vs. .42), though not significantly, $\chi^2(1) = 3.51$, $p = .061$, and had significantly higher confidence in the inserted item (3.14 vs. 2.37), $t(38) = 2.11$, $p = .041$, $d = 0.67$.

Discussion

Overall, participants in Experiment 2, like Experiment 1, demonstrated a limited ability to detect the manipulation of their external memory store. When asked to write down the words that had been presented, all participants wrote down the inserted word on the critical fourth trial, and participants expressed little knowledge explicitly that their external memory stores had been manipulated. The confidence data suggests that Experiment 2 participants were more sensitive to the manipulation than were Experiment 1 participants and again demonstrated that the inserted items and the items that had been presented previously were experienced (in terms of subjective

confidence) differently. The critical data from Experiment 2 was the recognition test performed at the end of the experiment without the participants having access to their external memory stores. Overall, participants appear to have integrated the words inserted into their external memory stores of the final lists into their biological memories for that list. Indeed, performance was statistically equivalent for inserted words and words that had actually been presented (though the inserted items were numerically lower on both measures). Importantly, inserted words were also more likely to be mistaken for words actually presented on list 4, than words from list 3 (though not after control for multiple comparisons), list 1 and new items.

Experiment 3

In Experiment 3, we sought to extend the results of Experiment 1 and 2 using a different test of whether individuals would accept the inserted word when they have access to their external memory store. As noted above, in the written recall protocol used in Experiments 1 and 2, participants wrote down all of the words but sometimes reported 0% confidence in the inserted words. This result is difficult to interpret. As noted above, it is possible that participants relied completely on the external store for deciding what to write (recall) but relied on their subjective experience when they reported confidence. Alternatively, this behavior might reflect a kind of demand characteristic or habit formed from the previous trials. Given the latter possibility, it would be beneficial to have converging evidence regarding participant's high likelihood of "accepting" the inserted word using a different measure during the initial memory test performed when they have access to their external store. As such, in Experiment 3 participants made the same confidence judgements they made in the final recognition task of Experiment 2 (i.e., 1-4) item-by-item (i.e., rather than by "copying" the words onto a sheet and providing a separate confidence judgement). Thus, the initial "memory" test is now recognition such that the participant, with access to their

external store, is presented with each word and the participant makes a 1-4 confidence judgement with respect to whether the word had been presented. The 1-4 confidence judgement also provides us a transparent measure of individual's subjective sense that the inserted word has been presented (i.e., definitely not, probably not, probably presented, definitely presented). We included a manipulation of location similar to Experiment 1 and included the recognition without external memory store task as in Experiment 2. Experiment 3 was pre-registered at https://osf.io/pu25h/.

Methods

Participants. One hundred and forty-seven students from the University of Waterloo participated for course credit. We planned to stop data collection once we had 120 (n = 60) useable participants (slightly more were collected as potential replacements if participants needed to be removed). This was based on having .80 power to detect at least a 25% difference (based on Experiment 1) in participants' ability to pick out the inserted word on the final question of the post-task questionnaire. Using the z test "Proportions: Difference between two independent proportions" in G*Power 3.1.9.2 (alpha = .05), this would require approximately 120 participants (Faul et al., 2007).

Apparatus. The apparatus was the same as in Experiment 3 with the exception that it was completely computer administered (as opposed to incorporating paper and pen methods) and was programmed to include the new recognition tasks on each trial.

Stimuli. The stimuli were largely the same as in Experiment 2, however, we changed the word lists so that (1) words that were often incorrectly encoded in previous experiments were replaced, (2) within list word order was not the same as in previous experiments, and (3) the inserted words were different. The mean word log frequency of the non-inserted/actually presented words was 7.58 (using frequency count from S*UBTLEX-UK*; Van Heuven, Mandera, Keuleers, &

Brysbaert, 2014) and mean word length was 6.46 letters. The inserted words (which were the same across participants) had a mean word log frequency of 7.56 and mean word length was 6.40 letters. These words did not differ in mean word length, $t(4.57) = 0.08$, $p = .940$, $d = 0.03$, or word frequency, $t(5.34) = .03$, $p = .979$, $d = .01$. The five lists were counterbalanced across participants, such that each list appeared on each trial (1 to 4) and acted as the unpresented list (used in the final recognition with external store task) equally often.

Recognition with External Memory Store. On each trial (four in total), we replaced the previously used recall task with a recognition task wherein participants were presented the word list one at a time on screen and rated whether it had been presented originally using the same 1 to 4 rating scale as used in the final recognition without external memory store phase in Experiment 2. For Experiment 3, endorsement and confidence during recognition refers to the % yes to being originally presented, and the 1 to 4 rating (as opposed to the previously used 0 to 100% confidence rating) for each word, respectively. The words presented on this task were the same as the words presented during encoding, and word presentation order was also matched. The only exception to this was in the case of the final trial. Specifically, the recognition task of the final trial presented the word that was inserted word (not originally presented) into their encoded (typed) list in its corresponding placement (i.e., if the word was inserted in the first position of their encoded list, the word also appeared as the first word during the final trial recognition task).

The *Recognition without External Memory Store* task was the same as in Experiment 2. The *Post-task questionnaire* used was identical to the one used in Experiments 1 and 2 with a slight modification such that participants answered "yes" or "no" to each of questions 1 to 3. Only when participants answered "yes" would they be asked to elaborate upon their answer. Question 4 remained the same as in Experiments 1 and 2.

Procedure. The procedure used in Experiment 3 was the same as in Experiment 2 with the exceptions of (1) the final list manipulation being the insertion of the word in either the first or the middle position and (2) participants completing a recognition task on each trial instead of the written recall tasks.

Results

Thirteen participants had to be replaced due to technical or protocol-based issues. There were 254 (3% of total) cases wherein participants did not encode a word properly into their external memory storage. Thus, these cases could not be included in endorsement and confidence analyses of the *Recognition with* and *without external store* tasks. There were three instances within the middle position condition wherein the researcher inserted the word into a position that was one later than protocol, however, the data from these participants were still included in analyses.

Participant answers to the post-task questionnaires were coded in the same manner as previously, except that participant answers were assigned a 0 for no knowledge and a 1 for at least some knowledge (referred to as "Some Knowledge" in Table 1; no intermediate level of 0.5). Two participants were replaced due to answering question 4 of the post-task questionnaire in a manner that demonstrated zero understanding. Initial agreement between coders was 95%. Disagreements were reconciled in the same manner as in the previous experiments.

Recognition with External Memory Store (Trial 4). Participant ratings were converted to yes/no answers (1 and 2 as "no"; 3 and 4 as "yes") to measure endorsement (i.e., that the item was originally presented). Participants with words inserted into the middle position of their list reported that the inserted word had been presented on the final list 78% [CI: 65%, 87%] of the time and did not differ significantly from participants with words inserted into the first position of their list (first: 77% [CI: 63%, 85%]), $\chi^2(1) = 0.05$, $p = .827$. As such, reported analyses comparing inserted

words to non-manipulated words are, hereon, collapsed across position condition. The analyses for each position separately revealed qualitatively similar results. Overall, the rate at which participants said yes to the actually presented words on list 4 was 98% [CI: 97%, 99%]. The rate at which participants said "yes" to the inserted word was statistically different from the rate at which participants said "yes" to words that were actually presented on the final list (final list non-manipulated: 98% [CI: 97%, 99%]), $t(119) = 5.66$, $p < .001$, $d = 0.52$, the rate at which participants said "yes" to all non-manipulated words, (all non-manipulated: 97% [CI: 97%, 98%]), $t(119) = 5.45$, $p < .001$, $d = 0.50$, and the rate at which participants said "yes" to the word adjacent (one after) the inserted word (one after: 98% [CI: 91%, 99%), $\chi^2(1) = 21.94$, $p < .001$ (this analysis was not pre-registered).

When confidence was treated as a continuous response, participants expressed high confidence that items inserted into their external memory had been presented originally. That is, in the middle position condition, mean confidence for inserted words was 3.35 [CI: 2.98, 3.60] and in the first position condition, mean confidence in the inserted word was 3.32 [CI: 2.95, 3.58]. Participants with words inserted into the first position of their list did not differ significantly in their confidence in the inserted word (as being presented originally) from the confidence of participants with words inserted into the middle position, $t(117.67) = 0.15$, $p = .882$, $d = 0.03$. This difference was also not significant when using a non-parametric test, $W = 1809$, $p = .952$. The high level of confidence overall in the inserted words was also apparent in the distribution of responses across the four options (see Table 3). That is, for both first and middle insertions the majority of the responses fell into the fourth response option "Word was definitely presented…"

Table 3. Mean percentage of the total number of responses for each confidence rating (1 to 4) for the inserted word on the recognition with external memory store (Trial 4) task of Experiment 3.

	Confidence Rating			
Condition	1: Definitely Not	2: Probably Not	3: Probably	4: Definitely
First	22	2	0	77
Middle	18	3	3	75

Note. Rows may not sum to 100 due to rounding.

Overall confidence for words that were actually presented on list 4 was 3.93 [CI: 3.89, 3.95]. Participants were significantly less confident that the inserted word (collapsed across first and middle conditions) had been presented than they were that the other final list (non-manipulated) words had been presented (final list non-manipulated: 3.93 [CI: 3.89, 3.95]), $t(119) = 5.65$, $p < .001$, $d = 0.52$, using a non-parametric test $V = 812$, $p < .001$, that the non-manipulated words across all lists had been presented (all non-manipulated: 3.91 [CI: 3.89, 3.92]), $t(119) = 5.44$, $p < .001$, $d = 0.50$, using a non-parametric test $V = 1798$, $p = .083$, and that the adjacent item (one after) to the inserted item had been presented (adjacent: 3.90 [CI: 3.77, 3.96]), $t(119) = 5.36$, $p < .001$, $d = 0.49$, and using a non-parametric test $V = 428$, $p < .001$.

Recognition without External Memory Store. As in Experiment 2, participants exhibited clear evidence that the manipulation of their external memory store had altered their biological memory for the final list in Experiment 3 (see Fig. 3). Participants with words inserted into the middle position reported that the inserted word had been presented on the final list 45% [CI: 32%, 57%] of the time and did not differ significantly from participants with words inserted into the first position (first: 50% [CI: 37%, 62%], $t(118) = 0.54$, $p = .587$, $d = 0.10$. A similar pattern emerged with confidence treated as continuous, (first: 2.45 [CI: 2.10, 2.78]; middle: 2.30 [CI: 1.97, 2.63]), $t(117.71) = 0.62$, $p = .537$, $d = 0.11$. As such, all analyses hereon are collapsed across position condition.

The rate at which participants said "yes" to the inserted word was statistically different from the rate at which participants said "yes" to words that were actually presented on the final

list (70% [CI: 67%, 73%]), $t(119) = 5.07$, $p < .001$, $d = 0.46$, but were higher than the rate at which participants said "yes" to words presented on list 3, (35% [CI: 31%, 38%]), $t(119) = 2.84$, $p = .005$, $d = 0.26$, list 1, and new words (list 1: 18% [CI: 14%, 22%] vs. new: 4% [CI: 2%, 7%]), all $ts \geq 6.60$, $ps \leq .001$, $ds \geq 0.60$.

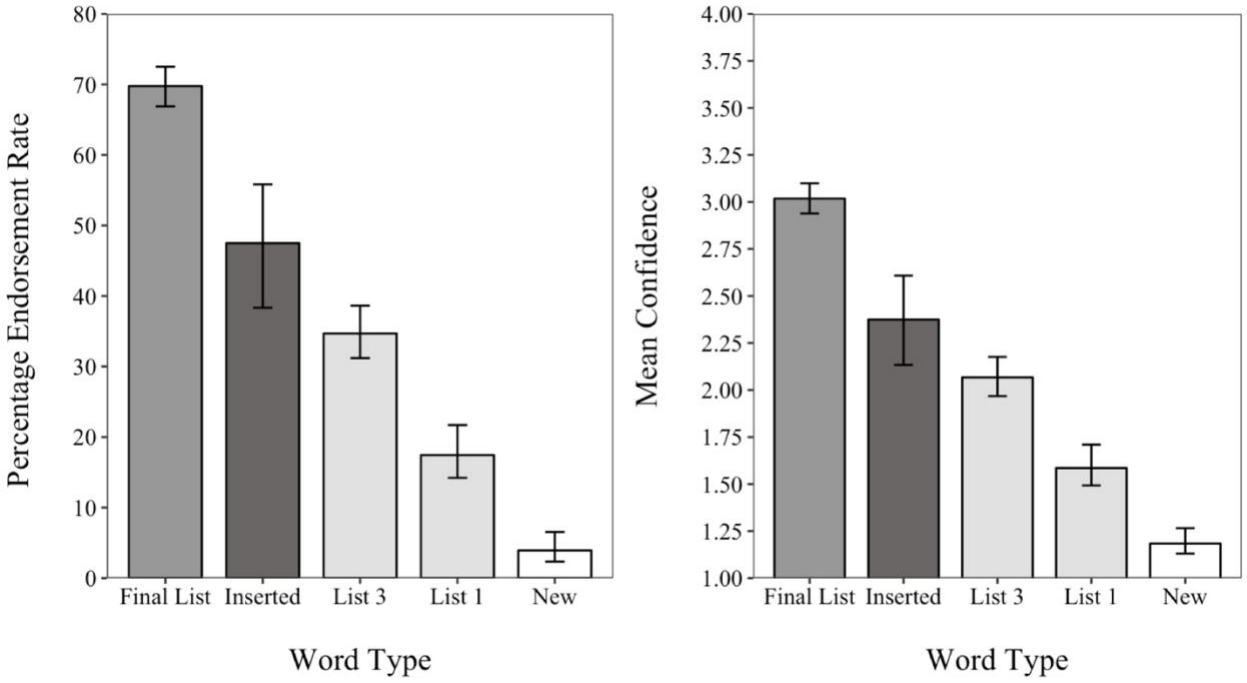

Figure 3. *Individual's responses in the final recognition test without the external memory store in Experiment 3 as a function of word type. Left Panel: Percentage of "Yes" responses where "3" and "4" are coded as "Yes." Right Panel: confidence responses treated as continuous (i.e., from 1 to 4). Error bars are bias corrected accelerated bootstrap confidence intervals using 10 000 replications.*

This same pattern emerged when participant responses were treated as continuous. That is, the confidence in the inserted word was less than actually presented final list words (inserted: 2.38 [CI: 2.13, 2.60] vs. list 4: 3.02 [CI: 2.94, 3.10]), $t(119) = 5.52$, $p < .001$, $d = 0.50$, and more than list 3 (list 3: 2.07 [CI: 1.97, 2.18]), list 1 (list 1: 1.59 [CI: 1.49, 1.71]) and new items (new: 1.18 [CI: 1.13, 1.27]), all $ts \geq 2.56$, $ps \leq .012$, $ds \geq 0.23$. Lastly, pairwise differences between non-manipulated word types (i.e., list 4, list 3, list 1, and new) were, again, all significant in the percentage of "yes" responses, all $ts \geq 7.54$, $ps < .001$, $ds \geq 0.69$, and when answers were treated

as continuous, all *t*s ≥ 8.05, *p*s < .001, *d*s ≥ 0.73. A qualitatively similar pattern emerged when using non-parametric paired comparisons. As in Experiment 2, a number of pairwise comparisons were reported in the above section. While all of these comparisons were pre-registered, it might be informative to consider the results when applying a correction for multiple comparisons. The only pairwise comparisons that were no longer significant were the comparisons between the inserted word and the list 3 words for endorsement, $t(119) = 2.84$, $p = .054$, $d = 0.26$, and confidence considered continuously, $t(119) = 2.56$, $p = .118$, $d = 0.23$. The latter analyses were not pre-registered. Ssee Table 4 for the percentage of each response (1-4) as a function of word type.

Table 4. Mean percentage of the total number of responses for each confidence rating (1 to 4) as a function of word type in the final recognition test without the external memory store in Experiment 3.

	Confidence Rating			
Word Type	1: Definitely Not	2: Probably Not	3: Probably	4: Definitely
Final List	19	11	19	51
Inserted	43	10	15	33
List 3	48	18	15	20
List 1	68	14	8	10
New	90	7	2	1

Note. Rows may not sum to 100 due to rounding.

Post-Task Questionnaire. As noted earlier, knowledge was coded in a binary manner. The first ten participants were not required (but were still asked) to pick out a word on question 4, however, the remaining participants were. Participants with words inserted into the first position of their final list demonstrated more knowledge of the inserted word on question 1 (first: 0.13 [CI: 0.05, 0.22]; middle: 0.03 [CI: 0, 0.08]), $\chi^2(1) = 3.93$, $p = .048$. Participants with words inserted into the first position of their list did not differ from participants with words inserted into the middle of their list for question 2 (first: 0.05 [CI: 0, 0.12]; middle: 0.00 [CI: 0, 0]), $\chi^2(1) = 3.08$, $p = .079$, question 3 (first: 0.05 [CI: 0, 0.10]; middle: 0.05 [CI: 0, 0.10]), $\chi^2(1) = 0$, $p = 1.00$, or for

question 4 (first: 0.27 [CI: 0.15, 0.37]; middle: 0.18 [CI: 0.08, 0.28]), $\chi^2(1) = 1.19$, $p = .274$. For question 4, participants were significantly above chance (~ 4.88%) in picking out the inserted word in both the middle position, $t(59) = 2.67$, $p = .010$, $d = 0.34$, and first position conditions $t(59) = 3.78$, $p < .001$, $d = 0.49$.

Unlike in previous experiments, we preregistered an extension of the questionnaire knowledge analyses. Specifically, we use chi-square analyses to investigate whether participants with words inserted into the first position of their lists differed from participants with words inserted into the middle positions of their lists in whether they answered "yes" or "no" to questions 1 to 3. Across all questions, there were no differences between the position conditions for answering yes or no, χ^2s ≤ 1.45, ps $\geq .228$. Table 1 presents the breakdown for knowledge type (0: none vs. 1: some) demonstrated by participants in each condition. As in previous experiments, participants expressed limited knowledge that a word had been inserted into their external store.

Exploratory Analyses

Response Time. We pre-registered an exploratory analysis of response time. For the Recognition with External Memory Store (Trial 4) task, we compared response time for all participants who responded "yes" (i.e., 3 or 4 on the confidence scale) to both the inserted item and a control item. The latter item differed across the first and middle position conditions. For participants with the word inserted into the first position, the control item was the first position item from participants in the middle position condition. For participants with the word inserted into the middle position, the control item was the middle position item from participants in the first position condition. Two Welch's two-sample t-tests were conducted on response times (in ms), and both revealed that participants responded to the inserted item slower. This difference was significant in the middle condition (control: 1092 [CI: 949, 1302]; inserted: 1559 [CI: 1265,

1966]), $t(67.38) = 2.34$, $p = .022$, $d = 0.49$, but not the first condition (control: 2670, [CI: 2240, 3433]; inserted: 2931 [CI: 2605, 3386]), $t(96.25) = 0.75$, $p = .454$, $d = 0.14$. A similar pattern emerged if log RTs were used but the middle condition comparison was marginally significant (control: 6.86 [CI: 6.73, 6.99]; inserted: 7.10 [CI: 6.90, 7.30]), $t(81.64) = 1.97$, $p = .052$, $d = 0.40$, whereas the first condition comparison was significant (control: 7.69 [CI: 7.56, 7.86]; inserted: 7.90 [CI: 7.79, 8.02]), $t(100.57) = 2.12$, $p = .037$, $d = 0.40$. Thus, there seems to be some evidence that inserted items are being processed differently than words that were actually presented during the encoding phase even when participants respond "Yes" to them.

Relation between Recognition with and without the list. As in Experiment 2 we conducted an exploratory analysis of the relation between responses to the inserted item (on trial 4) when participants had access to the external memory store and responses to the inserted item during the final recognition task completed without access to their external memory store. Participants who endorsed the inserted word when they had access to the external store (Trial 4) were significantly more likely to endorse the inserted word (endorsed on trial 4: .54; did not endorse on trial 4: .26), $\chi^2(1) = 5.43$, $p = .020$, and had higher confidence (when treated continuously) in the final recognition task completed without access to their external memory store (endorsed on trial 4: 2.54; did not endorse on trial 4: 1.81), $t(118) = 2.56$, $p = .012$, $d = .56$. The analyses in this section were not preregistered.

Discussion

Using a different memory task to assess whether, when performing the task with their external store, individuals would accept an inserted word, again revealed a surprising level of susceptibility to such manipulation. The majority of participants responded "yes" (i.e., 3 or 4 on the confidence scale) and the large majority of these "yes" response were done with high

confidence ("Word was definitely presented"). Again, participants showed almost a complete lack of "explicit" knowledge of the manipulation as assessed via the open-ended post-task questionnaire and had little success selecting the inserted word when asked. In the recognition without the external memory store task, as in Experiment 2, there was clear evidence that the inserted word was integrated into some participant's biological memories for that list. The rate at which participants responded "yes" (and confidence treated continuously) fell between words that had actually been presented (during the fourth trial) and words presented on list 3. While the ordering of the word types (i.e., inserted vs. list 3 vs. list 1 vs. new) remained the same as Experiment 2, participants in Experiment 3 could distinguish items actually presented from inserted items, overall. Experiment 3 also differed from Experiment 1 in that there was no clear effect of where in the list the inserted item was placed (first or middle position). While the means across the various measures of "noticing" that we employed suggested a higher level of noticing when the item was inserted into the first position (which was significant or marginal on a couple of these indices), it seems that if such an effect exists, then it is likely small and possibly sensitive to changes in context in which it is assessed (e.g., the differences between Experiment 1 and 3 in terms of measuring memory). The potential influence of location is worth further investigation in future studies.

General Discussion

Offloading to-be-remembered information onto external artefacts has long allowed us to escape the limitations of our biological memory (Clark, 2010a; Donald, 1991; Nestojko et al., 2013; Risko & Gilbert, 2016). Distributing memory demands across different locations, however, presents its own unique set of challenges. In the present investigation, we examined one such challenge: unlike our internal biological memories, externally storing to-be-remembered

information exposes it to direct manipulation. We outlined a number of cognitive factors, including a compromised biological memory for offloaded information (Eskritt & Ma, 2014; Sparrow et al., 2011) and metacognitions that would encourage reliance on the external store (Dunn et al., submitted; Risko & Dunn, 2015; Risko & Gilbert, 2016; Scorobia et al., 2007; Storm & Stone, 2015), that could make us susceptible to such manipulation. The results of Experiments 1 through 3 clearly support this idea. For example, in Experiment 3 almost 80% of participants thought that the inserted word had been presented (when performing their task with access to their external store) and most thought so with high confidence. In addition, across all of the post-task questionnaires, individuals rarely expressed any explicit knowledge that their external memories had been manipulated.

The vulnerability to external memory manipulation demonstrated here emerged even though the manipulation consisted of inserting a word that was unrelated to other items in the list[2], that individuals engaged in an effective encoding technique with the words that were actually presented (i.e., typing words into a file; Forrin et al., 2012), and that the duration between encoding and retrieval from the external store was exceptionally short (e.g., less than a minute). The latter is particularly important to note as it suggests that the susceptibility demonstrated here almost certainly underestimates what we would expect to observe in our day-to-day cognitive lives where days, months or even years might pass between encoding and retrieval from an external memory store. For example, there is some evidence that our sensitivity to memory distorting misinformation increases as the lag between original encoding and the presentation of the misinformation increases (Loftus, Miller & Burns, 1978).

In the above, we have suggested that the degree of vulnerability exhibited to the manipulation of our external stores was surprising. Experiment 3 provides an interesting means of

grounding this surprise, as we implemented the same type of memory report both when the individual had access to their external memory store and when they did not. In both cases, individuals were presented with "new" items (i.e., the inserted item on the final trial and the "new" items in the final recognition task). When a completely new word was added to their external store and they had access to that store, individuals responded "Yes" (a 3 or 4 on the confidence scale), that is they committed a false alarm, about 78% of the time. Comparatively, when a completely new word was presented in a task in which individuals did not have access to their external store, individuals said "Yes," also a false alarm, about 4% of the time. Thus, "new" words were rarely endorsed, except when those "new" words were inserted into individual's external memory store, in which case they were endorsed at a high rate (and with high confidence).

The results of Experiments 2 and 3 also demonstrated that our susceptibility to external memory manipulation has consequences for our biological memory for past events. Indeed, in Experiment 2, in a test of biological memory (i.e., a test of performance without access to the external store), individuals were unable to discriminate between words that had actually been presented previously and words that we had inserted into their external memory stores. In Experiment 3, individuals could discriminate between inserted and actually presented words, but still reported greater confidence that the inserted item had been presented than foils that had actually been presented on other lists. This result is interesting as it suggests that gaining access to one location (external store) within a distributed memory system can rather seamlessly provide access to other locations (internal, biological store).

One interesting perspective from which to view the present results is in the context of research on change blindness. Briefly, change blindness refers to the observation that, in many circumstances, individuals have difficulty detecting, for example, changes to scenes (Simons &

Levin, 1997). The present paradigm mimics those typically used to examine change blindness in that the original list is re-presented after a delay with a subtle change (i.e., the insertion of a new word) and we measure an individual's ability to detect that change. One critical difference in the current context is that participants are not explicitly looking for a change. Indeed, the paradigm here was designed to lead individuals to believe that their external memory stores are reliable in the sense of being unchanging (i.e., faithful representations of the information they had originally stored in there). In this manner, the present work resembles work on "choice blindness" (Hall, Johansson, Tarning, Sikstrom, & Deutgen, 2010; Johansson, Hall, Sikstrom & Olsson, 2005) and "memory blindness" (Cochran, Greenspan, Bogart & Loftus, 2016). The latter demonstration consisted of individuals viewing a crime and after a delay, reporting on aspects of that crime (e.g., selecting the perpetrator from a line up). After a subsequent delay, participants were presented with doctored versions of their memory reports (e.g., shown a different perpetrator under the cover that it was the one they had selected) and after another delay reported again on aspects of the crime (e.g., selecting the perpetrator from a line up). The doctored information was shown to influence the latter memory reports. The present research shares important features with this work, but it seems clear they differ in important ways as well. For example, the "false memory" rate here appears much higher (though it is difficult to directly compare across different measures) despite the fact that in Cochran et al. (2016) there were much longer delays between encoding and the initial memory report, between the latter and the doctored information, and between the latter and the final recognition task. In addition, in both "choice blindness" and "memory blindness", individuals make choices/memory reports not expecting to have to remember those choices/memory reports, whereas when we offload, we are typically doing so with the expectation that we will need to retrieve information from the external store in the future. Further investigation

aimed at understanding the similarities and differences between these phenomenon promises further insight into the malleability of our memory.

While our susceptibility to external memory manipulation is notable, the results of Experiments 1 to 3 also indicate that this susceptibility is not complete. For example, individuals consistently felt less confident that the items inserted into their external memories had been presented originally than they were that items that were actually presented when individuals had access to their external memory stores. Thus, it seems fair to conclude that we do not, as a whole, consider our external memories unquestioningly (though certainly some likely do). The exploratory response time analysis also provided an interesting perspective on this potential sensitivity to manipulation. That is, amongst individuals who responded 3 or 4 to both the inserted and a control item, individuals appeared to respond slower to the former than the latter.

Experiment 1 also demonstrated that individuals were more likely to detect a manipulation of their external memory when it affected the beginning rather than the middle serial position of their offloaded list. However, in a conceptual replication of this effect in Experiment 3, there was limited support for this. This might suggest that the mechanisms responsible for the primacy effect do not themselves influence the likelihood of detecting a manipulation of our external memories. A different interpretation, based on recent research (Kelly & Risko, in press), is that when individuals offload, information is sometimes encoded in such a manner so as not to produce a robust primacy effect (i.e., memorial benefit for initial items relative to intermediate items). The ambiguity with respect to the location manipulation or lack thereof suggests a need for future research to examine variables that might influence the likelihood that we detect a manipulation of our external memories. For example, provided the short time used here between encoding and

retrieval from the external store, it would be useful to examine longer delays (e.g., Loftus et al., 1978).

While there were clearly consistent overall patterns across the three experiments, there were also notable inconsistencies. For example, confidence (rated from 0-100%) in Experiment 1 and 2, for the comparable position (i.e., middle), was 79% in the former and 50% in the latter, and the endorsement rate for inserted items in the recognition test performed without the external store was 65% in Experiment 2 and 45% in Experiment 3 (for the comparable positions). In addition, Experiment 1 revealed effects of location of the insertion on a number of measures whereas this was not the case in Experiment 3. While these differences may reflect genuine differences resulting from variations in methodology across experiments (e.g., how memory reports were collected), it is important to also keep in mind that the insertion manipulation used here restricts us to a single observation per participant (and also as a result between subject designs). It seems reasonable that this might contribute to at least some of the noise in estimates across experiments. In addition, in order to facilitate insertion of the word into the participant's list (which had to be done online in a short amount of time) we fixed word order within lists. Across experiments we re-ordered and modified these lists, thus potentially interfering with what could have been idiosyncratic aspects of item order. Future work examining the possibility of using a paradigm featuring multiple manipulated trials using a within-participants design would be valuable as would amending the paradigm to better enable random assignment of items to locations within lists (this will be particular important for future investigations of the influence of item location).

Another interesting future direction will be to investigate how giving participants more control over what is offloaded might impact our susceptibility to manipulation of the external stores. In the present series of experiments, individuals were instructed to write down all of the to-

be-remembered words. This differs from contexts (arguably more typical) in which individuals might choose to offload or not on an item-by-item basis. In essence, this would ask individuals to solve both the "selection" problem and the "endorsement" problem (Arango-Munoz, 2013). When individuals can "choose" when to offload, individuals might exhibit a greater sensitivity to the susceptibility of their external memory to manipulation. That said, at least within a short-term memory task, individuals appear to rely on external stores extensively when it is available (Risko & Dunn, 2015) thus there might be limited opportunity to compare items selected for offloading versus not.

Conclusion. The ability to subvert our inherent cognitive limitations by offloading memory represents a critical tool in successfully navigating our complex cognitive lives. The present research highlights one of the inherent risks in this approach to "remembering." With the increasing availability of opportunities to offload memory, identifying and deeply understanding such risks will allow us to reap the most benefit out of our distributed memory systems.

Footnote

1. The difference in word frequency is difficult to interpret. The words that functioned as inserted words in Experiment 1 (and 2) all had different spellings in Canadian and American-English. This was a holdover from a previous experiment we had conducted and was not true in Experiment 3. Nevertheless, because of this issue we used British-English frequencies. If you instead use American-English frequencies (Brysbaert & New, 2009) with the spelling we used, then inserted words (5.73) are less frequent than non-inserted/actually presented words (6.15) they do not differ significantly, $t(7.53) = 1.06$, $p = .321$, $d = 0.21$. The same was true in Experiment 2. This issue was addressed in Experiment 3.

2. While the inserted word was not related to the list a meaningful manner (e.g., they did not belong to the same category;), it is important to point out that the inserted word did "fit" into the context of the list in that it consisted of unrelated words. We suspect that this fact likely increased the likelihood of accepting the inserted word relative to a context, for example, where the lists were related in some manner and the inserted word did not (e.g., inserting the word "apple" into a list of sleep related words). That said, we would also expect a rather high rate of acceptance if we had used related lists and inserted a word that fit in that context (e.g., inserting the word "sleep" into a list of sleep related words)

References

Arango-Muñoz, S. (2013). Scaffolded memory and metacognitive feelings. *Review of Philosophy and Psychology, 4*, 135-152.

Brown, G. D., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107*, 127-181.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 4: A semantic resource for sentiment analyses based on conceptual primitives. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2666-2677.

Carr, N. (2008). Is Google making us stupid? *The Atlantic, 302*, 56-62.

Clark, A. (2010a). *Supersizing the Mind*. Oxford, UK: Oxford University Press.

Clark, A. (2010b). Memento's revenge: The extended mind, extended. In R. Menary (Ed.), *The Extended Mind* (pp. 43-66). Cambridge, MA: MIT press.

Cherkaoui, M., & Gilbert, S. J. (2017). Strategic use of reminders in an 'intentional offloading' task: Do individual with autism spectrum conditions compensate for memory difficulties. *Neuropsychologia, 97*, 140-151.

Cochran, K. J., Greenspan, R. L., Bogart, D. F., & Loftus, E. F. (2016). Memory blindness: Altered memory reports lead to distortion in eyewitness memory. *Memory & Cognition, 44*, 717-726.

Donald, M. (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge, MA: Harvard University Press.

Eskritt, M., & Ma, S. (2014). Intentional forgetting: Note-taking as a naturalistic example. *Memory & Cognition. 42*, 237-246.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin and Review, 9*, 59-79.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Ferguson, A. M., McLean, D., & Risko, E. F. (2015). Answers at your fingertips: Access to the Internet influences willingness to answer questions. *Consciousness and Cognition, 37*, 91-102.

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40*, 1046-1055.

Gilbert, S. J. (2015a). Strategic offloading of delayed intentions into the external environment. *The Quarterly Journal of Experimental Psychology, 68*, 971-992.

Gilbert, S. J. (2015b). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of object memory ability. *The Quarterly Journal of Experimental Psychology, 68*, 245-260.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5*, 351-360.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition, 117*, 54-61.

Hothorn, T., & Hornik, L. 2019. exactRankTests: Exact distributions for rank and permutation tests R package version 0.8-30. Retrieved from https://CRAN.R-project.org/package=exactRankTests

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*, 116-119.

Kelly, M. O., & Risko, E. F. (in press). Offloading memory: Serial position effects. *Psychonomic Bulletin & Review.*

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*, 361-366.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 19-31.

Mazzoni, G. & Kirsch, I. (2002). Autobiographical memories and beliefs: A preliminary metacognitive model. In Perfect, T. J., & Schwartz, B. L. (Eds.), *Applied Metacognition* (pp. 121-145). Cambridge, UK: Cambridge University Press.

Nash, R. A., & Wade, K. A. (2009). Innocent but proven guilty: eliciting internalized false confessions using doctored-video evidence. *Applied Cognitive Psychology*, *23*, 624-637.

Nash, R. A., Wade, K. A., & Lindsay, D. S. (2009). Digitally manipulating memory: Effects of doctored videos and imagination in distorting beliefs and memories. *Memory & Cognition*, *37*, 414-424.

Nestojko, J. F., Finley, J. R., & Roediger, H. L. (2013). Extending cognition to external agents. *Psychological Inquiry, 24,* 321-325.

Pezdek, K., Finger, K., & Hodge, D. (1997). Planting false childhood memories: The role of event plausibility. *Psychological Science, 8,* 437-441.

Risko, E. F., & Dunn, T. L. (2015). Storing information in-the-world: Metacognition and cognitive offloading in a short-term memory task. *Consciousness and Cognition, 36,* 61-74.

Risko, E. F., & Gilbert, S. (2016) Cognitive offloading. *Trends in Cognitive Sciences, 20,* 676-688.

Risko, E. F., Ferguson, A. M., & McLean, D. (2016). On retrieving information from external knowledge stores: Feeling-of-findability, feeling-of-knowing and Internet search. *Computers in Human Behavior, 65,* 534-543.

Risko, E. F., Medimorec, S., Chisholm, J., & Kingstone, A. (2014). Rotating with rotated text: A natural behavior approach to investigating cognitive offloading. *Cognitive Science, 38,* 537-564.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology, 21,* 803-814.

Scorobia, A., Lynn, S. J., Hessen, J., & Fisico, S. (2007). So that's why I don't remember: normalizing forgetting of childhood events influences false autobiographical beliefs but not memories. *Memory, 15,* 801-813.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1,* 261-267.

Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences*, *9*, 521-560.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science. 333*, 776-778.

Sterelny, K. (2004) Externalism, epistemic artefacts, and the extended mind. In R. Schantz (Ed.), *The Externalist Challenge* (pp. 239-254). Berlin, DE: Walter de Gruyter.

Storm, B. C., & Stone, S. M. (2015). Saving-enhanced memory: The benefits of saving on the learning and remembering of new information. *Psychological Science*, *26*, 182-188.

Storm, B. C., Stone, S. M., & Benjamin, A. S. (2017). Using the Internet to access information inflates future use of the Internet to access other information. *Memory*, *25*, 717-723.

Strange, D., Gerrie, M. P., & Garry, M. (2005). A few seemingly harmless routes to a false memory. *Cognitive Processing*, *6*, 237-242.

Suprenant, M. & Neath, I. (2009). *Principles of Memory*. New York: Psychology Press.

Tousignant, J. P., Hall, D., & Loftus, E. F. (1986). Discrepancy detection and vulnerability to misleading postevent information. *Memory & Cognition*, *14*, 329-338.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176-1190.

Watson, J. M., Balota, D. A., & Roediger III, H. L. (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory and Language*, *49*, 95-118.

Ward, A. F. (2013). One with the cloud: Why people mistake the Internet's knowledge for their own (Unpublished doctoral dissertation). Harvard University, Cambridge, Massachusetts.

Wegner, D. M. (1995). A computer network model of human transactive memory. *Social Cognition, 13,* 319-339.