White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/187851/

Version: Accepted Version

Article:
Baddeley, AD, Atkinson, AL, Hitch, GJ et al. (2021) Detecting accelerated long-term forgetting: A problem and some solutions. Cortex, 142. pp. 237-251. ISSN: 0010-9452

https://doi.org/10.1016/j.cortex.2021.03.038

© 2021 Elsevier Ltd. All rights reserved. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/.

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Detecting Accelerated Long-term Forgetting: A problem and some solutions

Alan D. Baddeley[1*], Amy L. Atkinson[2], Graham J. Hitch[1] and Richard J. Allen[2]

[1]Department of Psychology, University of York, Heslington, York YO10 5DD

[2] School of Psychology, University of Leeds, Leeds, LS2 9JT

Author notes

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The data and materials are freely available on the Open Science Framework webpage: https://osf.io/4x23b/

*Correspondence should be addressed to: A.D. Baddeley, Department of Psychology, University of York, Heslington, York YO10 5DD, email: ab50@york.ac.uk.

MANUSCRIPT IN PRESS, CORTEX

Abstract

While many memory disorders occur with normal rates of forgetting, an accelerated rate of long-term forgetting (ALF) may occur, sometimes in the absence of a learning deficit. Detecting ALF presents a problem as it is desirable that the learned material is re-tested after each of several delays. This may result in earlier retrievals confounding later tests, with evidence suggesting that both positive and negative interaction can occur between successive tests. An earlier study (Baddeley et al., 2019) tested cued recall of a series of four crimes or four visual scenes by probing a different sample of features from all four crimes/scenes at each delay. Even though no question was asked twice, the interpolated tests markedly reduced the rate of forgetting. We suggest that this decelerated forgetting effect may result from the retrieval of probed features activating other associated features within that episode, hence facilitating their recall on subsequent tests. If so, the effect should be removed when only single and separate episodes, or individual items, are tested at each delay. We test this by probing a separate episode at each delay (Experiment 1), or by replacing integrated episodes with recognition memory for isolated words (Experiments 2 and 3) or visual scenes (Experiments 4 and 5). As predicted, we find no reduction in the rate of forgetting, in contrast to our earlier studies. The theoretical and clinical implications of our results are discussed. We conclude that the previously developed Crimes and Four Doors Tests (Baddeley et al., 2019) and the present single item recognition tests are complementary and are both likely to be necessary to ensure the reliable detection of ALF.

Keywords: Long-term forgetting, forgetting rate, retrieval practice, epilepsy, testing effect.

Detecting Accelerated Long-term Forgetting: A problem and some solutions

While many neuropsychological patients have problems in acquiring new information, once acquired such information is typically not forgotten at a more rapid rate than in healthy individuals, even when the episodic memory deficit is substantial as in the case of patient HM (Huppert & Piercy, 1977, 1978). Although claims of more rapid forgetting in patients suffering from Korsakoff Syndrome, Alzheimer's Disease, or indeed normal aging have been reported, the evidence was subtle and apparent only when tested by recall, in contrast to clear and marked deficits shown in episodic memory-based acquisition in these conditions (Cassel & Kopelman, 2019; Green & Kopelman 2002; Greene et al., 1996; Kopelman 1985; McKee & Squire 1992). Consequently, long-term forgetting is rarely assessed in neuropsychological memory testing with the long-term retention of novel material typically comprising delays of only 30 to 40 minutes with little monitoring over longer delays (Baker & Zeman, 2017; Elliott et al., 2014). Over time however, individual cases of apparently normal learning followed by dramatic forgetting began to appear (e.g. De Renzi & Lucchelli, 1993; Kapur et al., 1997; Martin et al., 1991).

Since that time, the field has expanded substantially. A review by Elliott et al. (2014) covers some 33 studies while Cassel et al. (2016) review 36 articles using a total of 100 different measures with 23 studies identified as methodologically sound. Having reviewed the overall evidence, both groups conclude that there is clear evidence for Accelerated Long-term Forgetting (ALF), notably in patients with temporal lobe epilepsy (TLE), with the most methodologically convincing examples based on patients with normal memory performance when tested both immediately and after short delays typically up to half an hour, but clear ALF when tested after longer

delays, such as one week or more. While a preserved capacity for initial learning is methodologically useful in establishing the phenomenon of ALF, there is no reason to believe that ALF is not also present in patients whose capacity for initial learning is impaired and for whom its detection may be harder but equally important. For that reason, tests are needed for which the establishment of an adequate level of initial learning is not too demanding.

We should make the additional point, however, that the term accelerated forgetting strictly refers to the shape of the forgetting function and for this, a demonstration of little loss over a short delay followed by later substantially greater forgetting is necessary. This therefore requires demonstration of relatively preserved retention followed later by a more rapid loss of information. This is typically studied by demonstrating normal forgetting over a brief delay, most typically 20 minutes. Our current design however concerns the impact of interpolated tests on the rate of forgetting in healthy young participants, a measure of overall speed rather than acceleration. As such it does not require the short delay over which little forgetting is to be expected. This study therefore forms part of an essentially pragmatic attempt to develop such measures, using a design based on our previous development of a sampled testing approach that appears to reduce rate of forgetting without leading to relearning (Baddeley et al., 2019). We wished to build on these results which therefore play a central role in the experimental designs that follow.

The need for novel test development is proposed by Elliott et al. (2014) who summarise some of the principal methodological issues raised and addressed in some earlier studies (Green & Kopelman, 2002; Isaac & Mayes, 1999; Kopelman &

Stanhope, 1997). Elliot et al. (2014) conclude with the need to systematically pilot a range of verbal and nonverbal tests and to identify those tests that offer the most reliable measure of ALF, bearing in mind the need for tests that are suitable for use in clinical practice. Elliott et al. also listed several other desirable characteristics for research in this area, namely that studies should use appropriately matched control groups, test both recall and recognition, avoid ceiling and floor effects, ensure an adequately long initial delay interval to avoid short-term recency effects, and equate initial learning across groups while avoiding over-learning. The present study forms part of an attempt to respond to this challenge and was initiated by an informal group of UK neurologists, neuropsychologists and cognitive psychologists who met with the aim of developing such tests.

A major problem that was not emphasised by Elliot et al. (2014) was the need to test the same patient several times. This is important theoretically, for instance, to establish whether ALF results from poor hippocampal-based consolidation operating over relatively short delays, or from the transfer of information from the hippocampus to other more durable cortical systems (Cassel & Kopelman, 2019; Mayes et al., 2019; Nadel & Moscovitch, 1997). It is also important clinically, as the pattern of ALF may potentially vary from patient to patient in both time of onset and rapidity of forgetting.

However, testing the same individual multiple times leads to the problem that each test is likely to influence subsequent tests, either maintaining performance by serving as a further rehearsal or potentially interfering with performance on later test trials. Within healthy individuals, many studies have now demonstrated that attempting to

retrieve material can, under certain circumstances, improve performance on a later memory test, and may be even more effective than a further learning trial (termed the 'retrieval practice effect'; see Roediger & Butler, 2011; Roediger & Karpicke, 2006 for reviews). Subsequent retrieval trials have also been suggested as a way of reducing forgetting in ALF patients (Jansari et al., 2010; Ricci et al., 2019*)*. As such, testing the same material multiple times may positively affect later performance in both patients with ALF and healthy controls. On the other hand however, there is extensive evidence that retrieving some items may negatively affect later memory for those that were not initially retrieved, an effect termed 'retrieval induced forgetting' (Anderson, 2003; Anderson et al., 1994; Macrae & Mcleod, 1999). It is also possible that early errors may be reinforced, thus affecting performance on the later tests both in healthy individuals (Kay, 1955) and more particularly in patients with memory impairment (Baddeley & Wilson, 1994; Kessels & de Haan, 2003).

In an attempt to develop material that was easy to learn while avoiding interference between passages, Baddeley et al. (2014) developed the Crimes Test in which four minor crimes are described, each comprising five features for example "The elderly Russian lady had her handbag snatched outside the cathedral by a young girl who then ran away". Retention is tested by probed recall for example "What was the nationality of the person robbed outside the cathedral?" which could also be asked in reverse order "Where was the crime committed against the Russian person?". Sampling equally across all four crimes yielded four sets of 20 questions with tests at each delay sampling a different set of questions, always selected from all of the four crimes. Initial results suggested levels of cued recall that avoided ceiling and floor effects for both young and older healthy participants. It was also found that broadly equivalent

recall scores were obtained when face-to-face learning was subsequently tested either by face-to-face or by telephone testing (Allen et al., 2019). The results proved sufficiently promising to add a visual version comprising four doors, each of which had five features, allowing questions such as "What was the object above the red door?", or "What was the colour of the door that had a statue above it?" (Baddeley et al., 2019).

However, recall performance at least up to one week remained relatively high, suggesting the possibility that even though no question was asked more than once, probing one feature of a crime might potentially activate the other features of the episode, hence enhancing subsequent recall. This possibility was investigated by Baddeley et al. (2019) who compared the performance of participants tested several times (immediately, after one day, one week and one month), with the performance of participants tested immediately and after one month with no intermediate testing. It proved to be the case for both the crimes and the doors tests that testing after an unfilled delay resulted in substantially more forgetting than when interpolated tests occurred after one day and one week (See Figures 1a and 1b). Broadly similar results were subsequently observed by Stamate et al. (2020) using passages based on four fables where testing of a subset of items from each fable at each delay resulted in decelerated forgetting both in young, healthy older participants and in patients with Alzheimer's disease.

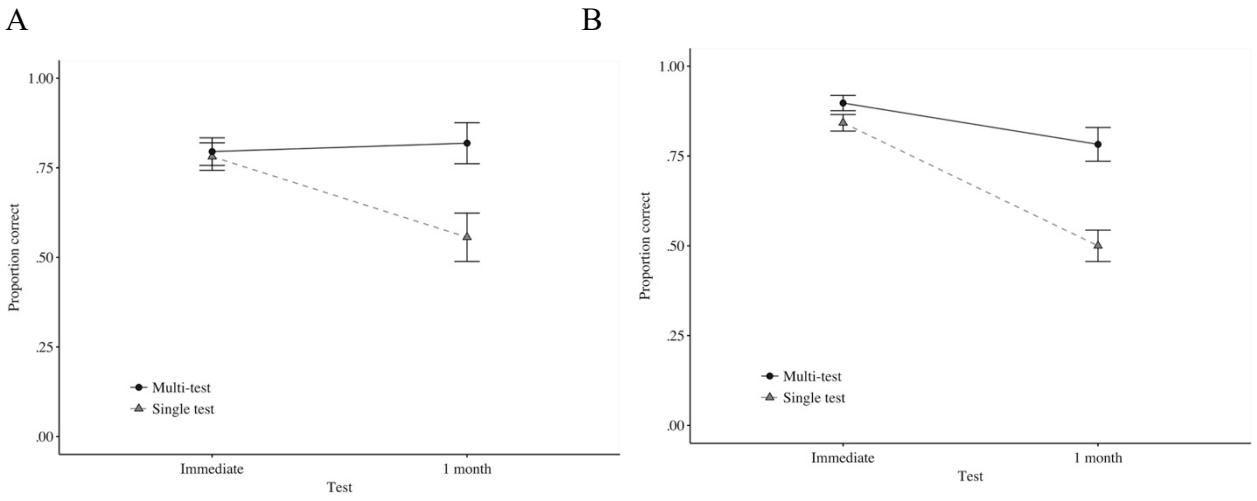

Figure 1A. Mean proportion correct (and SE) on the Crimes Test in the immediate and one-month tests as a function of test session and group. B displays equivalent data for the Four Doors Test. From Baddeley et al. (2019).

It is clearly the case that interpolated testing of a sample of features on each occasion is capable of delaying forgetting of other non-tested features. The multi-test groups in these studies thus had the advantage that the measures after different delays were clearly not independent since the earlier tests enhanced later recall. This in turn raised the question of whether the influence of one test on the others reflected new learning on each occasion or whether it was dependent on the priming of the original memory trace. This is an important distinction since patients with both the classic amnesic syndrome and deficits due to Korsakoff Syndrome or Alzheimer's Disease typically show preserved priming, while their capacity for new learning is clearly impaired (Graf et al., 1984; Tulving et al., 1982; Warrington & Weiskrantz, 1970). If priming is indeed preserved, then such patients should continue to benefit from the earlier tests, unlike TLE patients with ALF who show marked forgetting despite intervening tests (Drane, 2012; Laverick, 2018). We return to this issue in the discussion.

The positive effect of retrieval practice is typically reported when participants attempt to recall the whole initial learning task (Roediger & Butler, 2011; Roediger & Karpicke, 2006), while the equally well-established retrieval induced forgetting effect is typically based on single item retrieval (Anderson 2003; Anderson et al., 1994). The sampling procedure used in the Crimes and Four Doors Tests lies between these two; it involves learning a set of episodes or scenes but tests via the probed retrieval of individual features, with different features from each of the episodes tested at each delay. This is a potentially crucial factor according to the suggestion made by Baddeley et al. (2019) that with coherent prose material, probing one feature of an episode such as a crime or door scene may activate the whole episode. If so, cueing recall of features of each episode at each delay might serve to refresh the whole set of four crimes whether explicitly as part of a deliberate retrieval strategy or implicitly through spread of activation through the neural representation of the episode.

The series of experiments that follows attempts to study the effects of interpolated testing using material in which associations between successive tests are avoided. If, as proposed by Baddeley et al. (2019), the decelerated forgetting found using the Crimes and Four Doors tests results from the activation of the whole of each episode when a sample of features are cued, then we would expect the deceleration of forgetting to be lost when each test involves a single isolated episode or individual items.

Experiment 1

Experiment 1 aimed to test the hypothesis that the reduced rate of forgetting shown by Baddeley et al. (2019) and Stamate et al. (2020) stems from the tightly organised structure of the prose passages used. Experiment 1 adapted the original crimes material but tested a different crime at each delay in an attempt to avoid cross-delay priming effects. This is the method favoured by a number of earlier studies (e.g. Cassel et al., 2016; Jansari et al., 2010) who taught their participants a number of stories to a criterion before going on to test a separate story at each of a range of delays. They find clear evidence of substantial forgetting in both epilepsy patients and controls, although they did not include a condition for which intermediate tests were absent making it unclear whether the interpolated testing affected later retention either positively or negatively.

In the current experiment, two groups were used to examine this issue. One group was tested immediately, after 24 hours, and after one week, whilst the second group was tested only twice: immediately and after one week. If the capacity of intermediate tests to decelerate rate of forgetting is due to the cross-priming of material in our structured episodes (Baddeley et al., 2019; Stamate et al., 2020), then we might predict that testing a different story at each delay would eliminate this effect, resulting in either equivalent amounts of forgetting or possibly indeed greater forgetting due retroactive interference from the 24 hour test (Barnes & Underwood, 1959).

Method

Participants

All five experiments were performed by volunteer participants within the age range 18 to 30, principally undergraduate students, none of whom participated in more than one study.

Experiment 1 involved a total of 36 undergraduate or recent graduate participants from the University of York (aged between 18 and 30 years; mean age = 21.1 years; 17 female, 19 male). The study was approved by the University of York Psychology Department Ethics Committee and all participants gave informed consent. They were randomly allocated to two groups of 18, with testing shared by three experimenters. Test condition was alternated with experimenters, with half the experimenters testing the first and odd numbered participants in the multi-test condition and even numbered participants in the single-test condition, whilst the other half of testers followed the opposite pattern.

Materials

The material and data for this experiment and all other experiments within this series are freely available from the Open Science Framework (private link for review purposes only [to be replaced with a permanent link if accepted for publication]: (https://osf.io/649af/?view_only=3d55c01f964d46eeb6882e9829b8e73b). Our plan was to keep the material as similar as possible to that involved in the Crimes Test (Baddeley et al., 2014; 2019). However, when constructing the relevant passages it became clear that limiting probe questions to a single crime per test session resulted in severe constraints on the number of items that could be probed. This is best illustrated using two simple episodes, as for example in the phrase "The dog watched the bird" tested at one delay and "The cat ate the mouse" after a second delay. Given that one of the features has to be used to identify which episode is being tested, for

example the dog, questions are limited to the action "watched" and the object "bird". If however features can be sampled from either of the episodes, their potential number is clearly increased (e.g. "What creature did the watching?"; "What creature did the eating?"; "What happened to the bird?"; "What happened to the mouse?"; "What did the cat do?"; "What did the dog do?"). The advantage to the mixed sampling approach increases with the number of episodes potentially probed at any given delay. It also seems possible that mixing the episodes probed at any given delay may make the test more difficult than testing a single entire episode where the information contained in one question can be used to systematically build up the whole episode. Piloting the original Crimes material confirmed this problem, yielding a relatively limited number of questions and a very high level of performance. For that reason, each crime was elaborated by adding additional features, some containing potentially less memorable detail. Each crime therefore comprised six sentences that all had the same structure:

- Sentence 1: Nationality, gender and age of victim and location of crime.
- Sentence 2: Time of day, weather, what the victim felt.
- Sentence 3: The crime.
- Sentence 4: The age, gender and identity of the perpetrator, and the crime committed.
- Sentence 5: Clothing and height of the perpetrator.
- Sentence 6: Characteristics of a witness.

Design and procedure

A mixed design was used. Two groups were tested: one group completed tests immediately, 24 hours later and a week later (multi-test group) and the other group completed tests immediately and a week later (single-test group). The primary

research question concerned the amount of forgetting between the immediate test and the 28-day test in both groups. The experiment thus comprised a 2 (Group: multi-test vs single-test; between-subject) x 2 (Delay: immediate vs 28 days; within-subject) mixed design. The experimenter was not blind to the participant group membership, although participants were only informed of the number of tests they would need to complete after the encoding phase.

This required three separate crime episodes. Testing involved probing each crime, which was specified by its location. For example "For the crime committed at the bridge, what was the perpetrator's occupation?" This yielded a total of 14 questions for each crime as in the Crimes Test (Baddeley et al., 2019). Although the design required three crimes, piloting revealed a rather high level of performance when compared to the four crimes of the original test. We therefore added a fourth crime that was presented last and never tested, which served to prevent any recency advantage. We refer to this as an immediate test, although there is of course a delay interposed by the fourth untested crime. This serves to remove any undue advantage enjoyed by the final items, the recency effect that can complicate long-term forgetting rates by including a working memory component in the initial test (Greene et al., 1996). The recency component is avoided in our subsequent experiments by means of a brief interpolated visual or verbal task. The three to-be-tested crimes were presented and assessed in a counterbalanced order. As in the original studies, sentences were spoken at a steady rate with a two second pause after each sentence.

Following presentation of all four crimes, one of the first three was tested face-to-face by the experimenter asking the 14 probe questions with the participant responding

verbally. Regardless of performance only a single presentation occurred. Subsequent testing was conducted by telephone with the multi-test group being tested after 24 hours and a week, while the single-test group were tested only immediately and after a week. Across the experiment, the multi-test group were therefore tested on three of the crimes, whilst the single-test group were tested only on two. As with the study by Stamate et al. (2020) associations were probed in only one direction, from the location to the feature, hence no association was tested more than once. The study procedures and analyses for all experiments were not pre-registered.

Results

Figure 2A shows the mean percentage correct recall for the multi-test and single-test delay groups. A 2 (group: multi-test vs single-test; between-subject) x 2 (delay immediate vs one week; within-subject) mixed ANOVA was conducted on the proportion of questions answered correctly. There was a significant effect of delay ($F(1, 34) = 100.52$, p <. 001, $\eta_p^2 = .75$) but no significant main effect of group, indicating broadly equivalent levels of retention for the multi-test and single-test groups ($F(1, 34) = .79, p = .379, \eta_p^2 = 0.02$). No significant interaction was found between delay and group ($F(1, 34) = .45, p = .505, \eta_p^2 = .01$) indicating a similar rate of forgetting for the single and multi-test groups. Data suggest considerable variability between individuals, particularly at the 24-hour point, an important limitation if this test were to be used clinically (see Figure 2B and 2C).

Further analysis was conducted to examine the rates of forgetting across the delays in the multi-test group. There was a significant effect of delay ($F(2, 34) = 18.35, p < .001, \eta^2 = .52$), Bonferroni-Holm post-hoc comparisons revealed superior

performance on the immediate test relative to the 24 hour ($p < .001$) and 7 day ($p < .001$) tests. The difference between performance after 24 hours and 7 days approached significance ($p = .065$). For the single-test group, performance was significantly better on the immediate test than the 7 day test ($t(17) = 10.26$, $p < .001$).

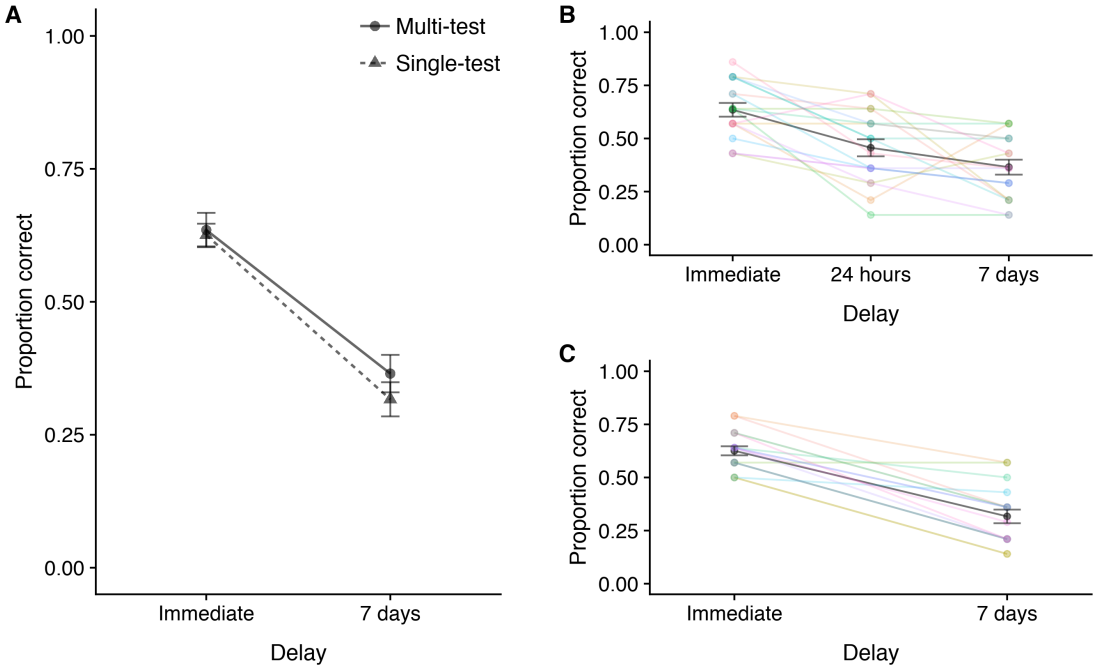

Figure 2. Panel A displays the mean proportion correct in Experiment 1. Panel B presents forgetting rates for each participant in the multi-test condition, whilst Panel C displays the equivalent data for the single-test condition. The dark line in Panels B and C reflects mean performance across participants. Error bars in all panels denote SE.

Discussion

The results of Experiment 1 are clear in suggesting that testing separate crimes at each delay removes the facilitating effect of intermediate testing, as is the case when features are sampled from each crime at each delay. Figure 2 suggests that we successfully avoided ceiling and floor effects. However, if this test were to be used clinically, it would likely require a higher level of initial learning, potentially resulting

in ceiling effects. Conversely, extending the test beyond one week would risk floor effects. These limitations encouraged us to move to material where items to be remembered were more clearly independent of strong associative links and where it would prove easier to control task difficulty. We opted to test recognition memory for individual words and scenes, using list length to control difficulty. This has the additional advantage of potentially adding a recognition measure to those of free or cued recall that are more frequently used (Cassel & Kopelman, 2019). This would also satisfy one of the criteria outlined by Elliott et al. (2014) for a desirable ALF assessment tool to include both recall and recognition.

Experiment 2

This study departs from the use of text, using instead, recognition memory for lists of words. Verbal recognition memory tends to be easier than recall (Mandler et al., 1969), which allows a larger sample of items to be included, potentially increasing overall test reliability. Performance will however be limited by guessing rate. A simple two-alternative Yes-No recognition would have a 50% guessing rate, thus requiring lengthy lists for adequate sensitivity. We therefore opted for four-alternative testing with each target presented together with three novel lures. Another complication concerned the selection of words. Here we opted for concrete nouns as is typically the case in memory studies. While relative ease of learning is advantageous, particularly in a clinical context, selecting random negative items or lures is likely to result in potential ceiling effects. We therefore chose to use target items from various semantic categories (e.g. furniture, animals), in each case testing them with three lures taken from the same category. Encoding strategy is a further potential source of variability. We therefore required each target word to be judged as

"pleasant" or "unpleasant", a task that participants find relatively easy (Warrington, 1984) but enhances performance (Baddeley & Hitch, 2017; Evans & Baddeley, 2018). Inter-word priming effects were minimised by ensuring that each category was tested only once at each of the four potential delay points. As we anticipated higher accuracy overall due to the use of recognition (Mandler et al., 1969), we included a 28-day test. If sensitive performance levels are still observed at the 28-day test, this may allow us to assess ALF for a longer period of time, thus potentially allowing us to detect ALF emerging more than one week after encoding. This would also be in line with some of our previous work, which has investigated delays at one month (Baddeley et al., 2019; Stamate et al., 2020). Given that performance was tested over a longer delay than in Experiment 1, an encoding phase criterion was used, with participants shown the list of words again if they failed to answer at least 75% (18/24) of the questions correctly on the immediate test. This is similar to that used in previous research (Baddeley et al., 2019; Stamate et al., 2020).

As in Experiment 1, our principal question concerned the amount of forgetting between the immediate test and the final test (28-day test) in the multi-test and single-test groups. If the reduced forgetting from interpolated tests shown by Baddeley et al. (2019) and Stamate et al. (2020) is due to inter-item priming, then the difference should be lost when inter-item associations are avoided, as was observed when only one crime was assessed at each delay in Experiment 1.

Method

Participants

A total of 32 healthy young adults (aged between 18 and 30 years) were randomly assigned to either a multi-test or a single test groups of 16 participants. As in Baddeley et al. (2019), multiple testers were employed, as would be the case were the tests to be used clinically, with each of eight testers testing a balanced sample of four participants. Ethical approval was granted by the University of York Psychology Department Ethics Committee. Neither participants nor testers had taken part in Experiment 1.

Materials

There were 24 semantic categories (e.g. fruit, disease), with 16 words in each category (384 words in total). Words were taken from the Battig and Montague (1969) norms, which contain lists of words that are most frequently associated with a particular category. Categories were selected if they contained at least 16 items. If the category contained more than 16 words, the most frequently identified words were selected. An item was replaced with the next available word if it contained two words (e.g. *can opener*), was a homonym (e.g. *chest*) or was considered to be more associated with a different category (e.g. *piano* appeared in the furniture category but may be more associated with musical instruments). Words were also replaced if they were an obvious synonym of another word in the set (e.g. *home* when *house* was already included in the set), were over 12 letters long (e.g. *chrysanthemum*), were deemed to be American English (e.g. *ladybug*) or were no longer considered to be frequently occurring (e.g. *mononucleosis*).

Follow-up tests. The 16 words within each category were split into four subsets of four words based on their frequency (Battig & Montague, 1969). The words within

each frequency-based subset were numbered from 1-4 (e.g. for Animals, *1. Cat - 2. Dog – 3. Horse - 4. Cow*) and a target item (e.g. *Horse*) selected at random. Four lists were then created, with each comprising 24 subsets. Each of the four lists contained one subset from each category (e.g. fruit). The order of subsets within each list was randomised. The four lists formed the 24 item tests to be administered at the various delays. One item within each subset was presented during the encoding phase, whilst the three other items served as lures.

Target items. The list of words to-be-presented during the encoding phase was created by pseudo-randomly selecting one word from each of the 96 subsets, such that each option (1, 2, 3, 4) was equally likely to be correct across the entire experiment. The to-be-presented items were the same for each participant.

Design and Procedure

The design used was the same as Experiment 1, except that in the multi-test group, participants completed four rather than three tests (immediately after encoding, then after 24 hours, 7 days and 28 days). Participants in the single-test group, completed only one delayed test, after 28 days. Given that there were four delay periods in this experiment, all four of the lists created were used in the multi-test group. Only two of the lists were used in the single-test group.

During the encoding phase, each word was visually presented to the participant for two seconds and read aloud by the experimenter. Participants were then given two seconds to decide whether the meaning of each word was pleasant or unpleasant before the next word was displayed and read aloud. They provided this pleasantness

judgement using a computer keyboard (pressing the q key for pleasant and p key for unpleasant). Before the start of the encoding phase, participants completed four practice trials to familiarise themselves with the task. The words used in the practice phase were the same for all participants, and did not belong to any of the 24 semantic categories used in the experiment. The order of words in the presentation phase were the same for all participants.

Following the encoding phase, participants completed a visual Spot the Difference task for one minute. This involved noting differences between two highly similar pictures and was implemented in order to reduce any potential recency effects (Baddeley et al., 2019; Elliott et al., 2014; Greene et al., 1996). Next, participants completed the immediate test, using one of the four test lists. The list order was fully counterbalanced across participants. During these tests, participants were read aloud the four words in each subset and asked to identify which word had been presented during the encoding phase. Chance rate was therefore 25%. If participants scored less than 75% on the immediate test, the encoding phase and the Spot the Difference task were repeated after which participants were retested using the same test material. If necessary, this process was repeated until participants reached the 75% criterion or for a maximum of three times.

The follow-up tests were conducted by telephone after 24 hours, 7 days, and 28 days (multi-test group) or after 28 days (single test group).

Results

Mean proportion correct and standard error (SE) are displayed in Figure 3A as a function of group and delay, whilst individual performance is displayed in Figures 3B and 3C. A 2 (Group: multi-test vs single-test; between-subject); x 2 (Delay: immediate vs 28 days; within-subject) mixed ANOVA was conducted. A significant main effect of delay emerged ($F(1, 30) = 395.67$, $MSE = .01$, $p < .001$, $\eta_p^2 = .93$), with participants exhibiting higher accuracy in the immediate test relative to the 28-day test. There was also a main effect of group ($F(1, 30) = 6.77$, $MSE < .01$, $p = .014$, $\eta_p^2 = .181.80$), with participants in the multi-test group showing higher overall accuracy than participants in the single-test group. However, no interaction between group and delay emerged ($F(1, 30) = 1.59$, $MSE = .01$, $p = .218$, $\eta_p^2 = .05$).

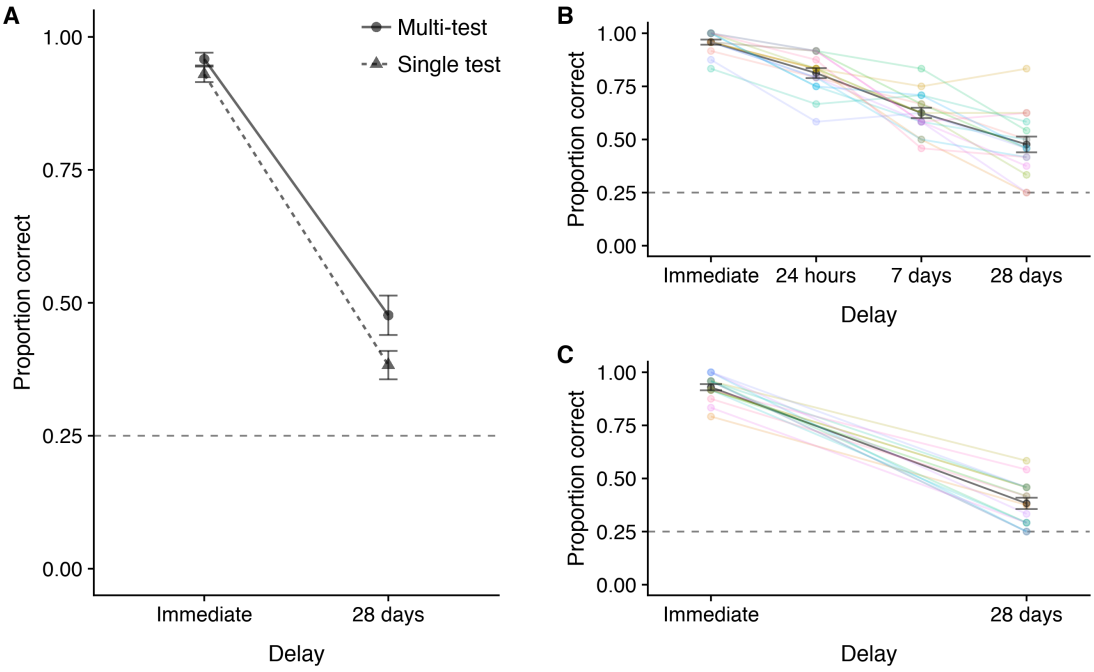

Figure 3. Proportion correct (and SE) in Experiment 2, as a function of delay and group. Figure 3A displays mean proportion correct as a function of delay and group, whilst 3B and 3C present individual participants' forgetting rates across the delays in the multi-test (B) and single-test groups (C). The darker lines reflect mean

performance and the horizontal lines at 0.25 reflects chance guessing rate. Error bars display SE.

Analysis was also conducted to investigate whether performance differed across the four delays in the multi-test and the two delays in the single-test group respectively. A repeated-measures one-way ANOVA with four levels (immediate, 24 hours, 7 days, 28 days) was conducted to explore this in the multi-test group. This revealed a significant effect of delay ($F(1.85, 27.81) = 77.23$, $MSE = .02$, $p < .001$, $\eta_p^2 = .84$). Bonferroni-Holm post-hoc comparisons revealed higher accuracy on the immediate test relative to the 24 hour test ($p < .001$), 7 days ($p < .001$) and 28 days ($p < .001$). The proportion of questions answered correctly was also higher on the 24 hours test, relative to the 7 day ($p < .001$) and 28-day tests ($p < .001$), and higher on the 7 day relative to the 28 day tests ($p < .001$). To investigate whether performance in the single-test group differed in the immediate and 28-day tests, a paired sample t-test was conducted. A significant difference emerged ($t(15) = 17.58$, $p < .001$, $d = 4.40$), with superior performance on the immediate test.

Discussion

In line with Experiment 1, the intervening tests had no reliable effect on rate of long-term forgetting. This is in contrast to the clear reduction of forgetting rate found in previous studies based on integrated prose or scenes (Baddeley et al., 2019; Stamate et al., 2020). The analysis indicated a main effect of group, with participants in the multi-test group exhibiting higher accuracy overall than the single-test group, presumably reflecting a chance difference in learning capacity between the groups. The interaction between group and delay, which might indicate reduced forgetting in

the multi-test group, did not reach significance. However, the chance difference between the two groups in overall learning capacity, together with the observation that individuals from both groups performed at, or near, chance levels at the longest delay suggests that the absence of a significant interaction should be viewed with caution. For that reason we chose to replicate this study, using a maximum delay of one week (as in Experiment 1), which Figure 3 suggests is sufficient to show clear forgetting while avoiding floor effects.

Experiment 3

Experiment 3 aimed to replicate the findings of Experiment 2, but with a reduced delay between the immediate test and final test (7 days, instead of 28 days used in Experiment 2). It was predicted that no interaction would emerge between group and delay, in line with the findings from Experiments 1 and 2.

Method

Participants

Forty-eight healthy young adults (age between 18 and 30 years) were randomly assigned to a multi-test or single-test group. There were 24 participants in each group. We again employed multiple testers with eight testers each testing a balanced sample of six participants. Neither testers nor participants had taken part in either of the previous experiments. The University of York Psychology Department Ethics Committee gave ethical approval.

Design, Materials and Procedure

The design, materials, and procedure were identical to Experiment 2, except that the 28-day test was omitted. Participants in the single-test group therefore completed tests immediately after encoding and after 7 days, whilst participants in the multi-test group completed tests immediately after encoding, after 24 hours, and after 7 days. The number of test lists used was therefore reduced to three, with the fourth test list being removed. The presentation phase was identical to that in Experiment 2, containing all 96 words from 24 distinct semantic categories.

Results

Mean proportion correct (and standard error (SE) is displayed in Figure 4A as a function of group and delay, whilst individual performance is displayed in Figure 4B and 4C. A 2 (Group: multi-test vs single-test; between-subject) x 2 (Delay: immediate vs 7 days; within-subject) mixed ANOVA was conducted. A significant main effect of delay emerged ($F(1, 46) = 334.62$, $MSE = .01$, $p < .001$, $\eta_p^2 = .88$), with participants exhibiting higher accuracy in the immediate test relative to the 7 day test. There was no main effect of group ($F(1, 46) = 2.51$, $MSE = .02$, $p = .120$, $\eta_p^2 = .05$), and no interaction between group and delay ($F(1, 46) < .01$, $MSE = .01$, $p = .966$, $\eta_p^2 < .01$).

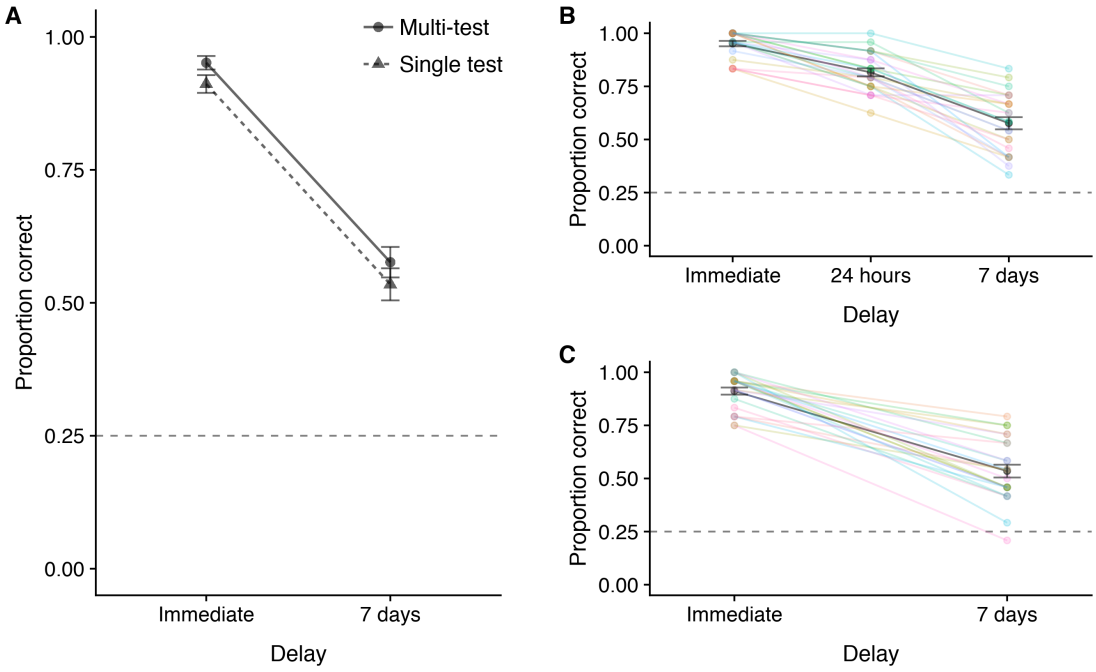

Figure 4. Mean proportion correct and SE in Experiment 3 as a function of delay and 4A. shows overall group data, 4B shows individual multi-test and 4B single test data. The darker lines in Panels B and C display the mean proportion correct across participants, whilst the dotted horizontal lines at 0.25 reflect chance guessing rate. Error bars denote SE.

Analysis was also conducted to investigate whether performance differed across the two delays in the multi-test and the one delay in the single-test group. A repeated-measures one-way ANOVA with three levels (immediate, 24 hours, 7 days) was conducted to explore this in the multi-test group. This revealed a significant effect of delay ($F(1.49, 34.24) = 132.04$, $MSE < .01$, $p < .001$, $\eta_p^2 = .85$). Bonferroni-Holm post-hoc comparisons revealed superior performance in the immediate test relative to the 24 hours test ($p < .001$) and the 7 days test ($p < .001$). Accuracy was also higher on the 24 hour test than the 7 day test ($p < .001$). To investigate whether performance in the single-test group differed in the immediate and 7 days tests, a paired sample t-

test was conducted. A significant difference emerged, ($t(23) = 12.20$, $p < .001$), with performance higher in the immediate test.

Discussion

The results of Experiment 3 replicate those of Experiment 2, except that the overall difference in performance between the two groups was no longer observed. Any suggestion of an interaction between group and delay was absent, implying that intervening tests had no influence on long-term forgetting under these conditions. Thus, Experiments 1-3 clearly show no facilitation effect from interpolating tests when individual episodes are tested or when isolated items are presented at encoding in verbal memory. This contrasts with the very substantial effects shown when integrated material allows inter-item priming across successive tests (Baddeley et al., 2019, Stamate et al., 2020) as shown in Figure 1.

As Elliott et al. (2014) suggested, it would be valuable if tests for ALF could be developed for both verbal and visual material. This was implemented in the case of cued recall through the Four Doors Test (Baddeley et al., 2019), which was also found to show very marked reduction of rate of forgetting from interpolated testing when features of each door were sampled at each delay. This again suggested that each door served as an integrated scene in which probing one feature primed or evoked the whole scene. We therefore extended our investigation in two further experiments using the same design as Experiments 2 and 3 but using complex visual scenes, again attempting to minimise within-test associations.

Experiment 4

Experiment 4 examined the effects of interpolating tests for visual scenes over a 28-day period. This experiment was therefore identical to Experiment 2, except that the materials were visual in nature instead of verbal.

Design

A 2 (Group: multi-test vs single-test; between-subject) x 2 (Delay: Immediate vs 28 days; within-subject) mixed design was used. The multi-test group were tested immediately, after 24 hours, 7 days, and 28 days, whilst the single-test group were tested immediately and after 28 days.

Participants

Thirty-two healthy young adults aged between 18 and 30 years completed the experiment. None had participated in any of the previous experiments. Ethical approval for this experiment was granted by the School of Psychology Ethics Committee at the University of Leeds. Sixteen participants were randomly assigned to the multi-test group, whilst the other sixteen were randomly assigned to the single-test group. There were five testers.

Materials

As in Experiments 2 and 3, there were 24 distinct categories of scenes, with each category comprising a different type of scene (e.g. skyscraper, fountain, bathroom, restaurant etc). Indoor and outdoor scenes were used, as opposed to everyday objects, to reduce verbal labelling and rehearsal as much as possible. Half of the categories were outdoor scenes (e.g. skyscraper, fountain) and half were indoor scenes (e.g. bathroom, restaurant). There were 16 images within each category. The images were

taken from the public image dataset curated by Computational Visual Cognition Laboratory MIT (http://cvl.mit.edu). The 16 scenes were randomly split into four subsets of four scenes. Four test lists were then created to be administered at the four delays. These four test lists each comprised 24 subsets (one subset from each of the 24 categories). The order of subsets within the test lists was randomised.

One scene within each subset was presented during the encoding phase and the three other scenes acted as lures. The scenes presented during the encoding phase were selected pseudo-randomly, such that each of the four options (1, 2, 3, 4) were correct a similar number of times across the experiment. The items presented during the encoding phase were the same for all participants.

Design and Procedure

The design and procedure were identical to Experiment 2, except for a few minor differences. Based on piloting, each image was presented on screen for 3 seconds in total, during which time participants had to provide a pleasantness judgement which was verbally stated and recorded by the experimenter. The interpolated task administered between encoding and the immediate test to reduce recency involved participants finding as many words as possible from the word "hippopotamus" (Baddeley et al., 2019). Finally, the follow-up tests (the 24 hours, 7 day and 28-day tests in the multi-test group, and the 28-day test in the single-test group) were completed online as the stimuli were visual in nature, thus preventing spoken telephone testing.

Results

Mean proportion correct is displayed in Figure 5A as a function of group and delay, whilst individual performance is displayed in Figure 5B and 5C. A 2 (Group: multi-test vs single-test; between-subject) x 2 (Delay: immediate vs 28 days; within-subject) mixed ANOVA was conducted. There was a significant main effect of delay ($F(1, 30) = 381.94$, $MSE = .01$, $p < .001$, $\eta_p^2 = .93$), with participants exhibiting higher accuracy in the immediate test relative to the 28 day test. However, no main effect of group emerged ($F(1, 30) = .10$, $MSE = .01$, $p = .760$, $\eta_p^2 < .01$), and there was no interaction between group and delay $F(1, 30) = .09$, $MSE = .01$, $p = .769$, $\eta_p^2 < .01$).

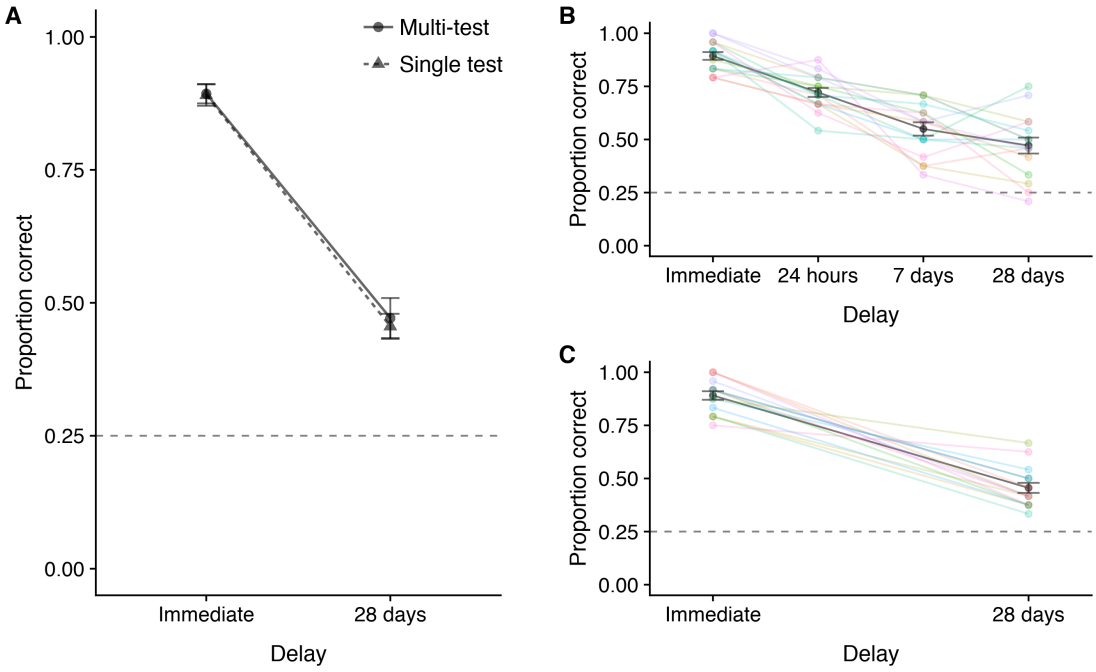

Figure 5. Mean proportion correct and SE in Experiment 4 as a function of delay and group. (5A). Figures 5B and 5C present proportion correct for individual participants in the multi-test (B) and single-test groups (C). The dark lines reflect individual performance and the dotted horizontal lines reflect chance guessing rate. Error bars denote SE.

Analysis was also conducted to examine whether the proportion correct differed across the four delays in the multi-test group and the two delays in the single test group. In the multi-test group, a repeated-measures one-way ANOVA was conducted with four levels (immediate, 24 hours, 7 days, 28 days). This revealed a significant effect of delay ($F(3, 45) = 54.45$, $MSE = .01$, $p < .001$, $\eta_p^2 = .78$). Bonferroni-Holm post-hoc comparisons revealed significantly higher accuracy in the immediate test than at 24 hours ($p < .001$), 7 days ($p < .001$), and 28 days ($p < .001$). There was also significant better performance in the 24 hours test relative to the 7 days ($p < .001$) and 28 days ($p < .001$) tests. The difference between performance after 7 days and 28 days approached significance ($p = .084$). To examine whether a significant effect of delay was observed in the single test group, a repeated measures t-test was performed. This revealed a significant difference between performance at the immediate test and the 28 day test ($t(15) = 13.21$, $p < .001$, $d = 3.30$), with higher accuracy in the immediate test.

Discussion

The results of Experiment 4 resemble those from the three earlier experiments in suggesting very clear evidence for forgetting, but no interaction between group and delay. This provides further evidence that the rate of forgetting over a one-month delay is not influenced by the presence of interpolated tests when non-integrated materials are used. Once again, however, there is a suggestion that the performance of some participants may possibly be constrained by floor effects at the 28-day test. We therefore conducted a fifth experiment which, like Experiment 3 used a final delay of 7 days. We thus compare two groups, a multi-test group tested immediately, after 24 hours and after a week, and a second group tested immediately and after 7 days.

Experiment 5

Experiment 5 aimed to examine the effect of interpolating tests on visual scenes over a period of 7 days.

Method

Participants

Thirty-two healthy young adults aged between 18 and 30 years participated. None had completed any of the previous experiments. Sixteen participants were randomly assigned to each group (multi-test group and single-test groups) and were tested by five testers. Ethical approval was granted by the School of Psychology Ethics Committee at the University of Leeds.

Design, Materials and Procedure

The design, materials, and procedure were identical to Experiment 4 except that the 28-day test was omitted. Participants in the single-test group completed tests immediately after encoding and after 7 days. Participants in the multi-test group completed tests immediately after encoding, after 24 hours, and after 7 days. The number of items presented was the same as Experiment 4 but only three test lists were used. The order of lists used was counterbalanced. The presentation phase was identical to Experiment 4, with all 96 images presented.

Results

Mean proportion correct and standard error (SE)) are displayed in Figure 6A as a function of group and delay. Performance for individual participants is displayed in Figure 6B and 6C. A 2 (Group: multi-test vs single test; between-subject) x 2 (Delay:

immediate vs 7 days; within-subject) mixed ANOVA revealed a significant main effect of delay ($F(1, 30) = 152.94$, $MSE = .01$, $p < .001$, $\eta_p^2 = .84$), with participants exhibiting higher accuracy in the immediate test relative to the 7 day test. There was no main effect of group ($F(1, 30) = .92$, $MSE = .02$, $p = .345$, $\eta_p^2 = .03$), although a significant interaction between group and delay was observed ($F(1, 30) = 5.25$, $MSE = .01$, $p = .029$, $\eta_p^2 = .15$). However, as shown in Figure 6A this suggests greater forgetting in the multi-test group, the opposite to the substantial effect found in multi-test groups with integrated material (Baddeley et al., 2019; Stamate et al., 2020), to the direction of the non-significant trend shown in Experiment 2 and the clear absence of any such effect shown in the other experiments in this series.

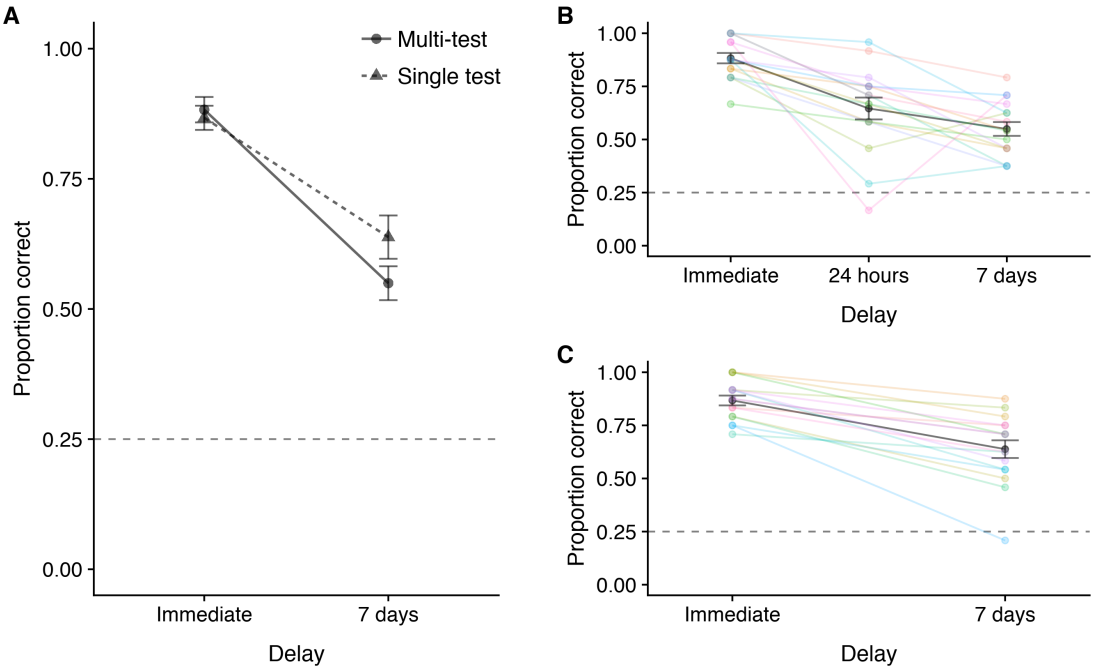

Figure 6. Mean proportion correct and SE in Experiment 5 as a function of delay and group. 6A shows overall group data while 6B and 6C present individual participants' scores. The darker line in Panels B and C display mean proportion correct. The error bars denote SE, whilst the horizontal lines at 0.25 present chance guessing rate.

To explore the significant interaction, Bonferroni-Holm corrected independent samples t-tests were performed to examine whether there was an effect of group at the different tests. This revealed no significant effect of group at either the immediate test ($t(30) = .46, p = .646, d = .16$) or the 7 days test ($t(30) = -1.68, p = .208, d = -.59$). Accuracy was significantly lower at the 7 day test relative to the immediate test in both the multi-test ($t(15) = 10.12, p < .001, d = 2.53$) and single-test ($t(15) = 7.31, p < .001, d = 1.83$) groups.

A one-way repeated measures ANOVA was conducted to explore whether there was a main effect of delay across the four conditions in the multi-test group. A significant effect of delay was observed ($F(2, 30) = 26.74, p < .001, \eta_p^2 = .64$). Bonferroni-Holm post-hoc tests revealed significantly better performance in the immediate test relative to the 24-hour test ($p < .001$) and the 7-day test ($p < .001$). There were no significant differences between the 24 hour test and the 7 day test ($p = .099$). To examine the effect of delay in the single test group, a repeated-measures t-test was conducted, revealing significantly better performance in the immediate test relative to the 7-day test ($t(15) = 7.31, p < .001, d = 1.83$).

Discussion

Experiment 5 again used visual scenes and once more found clear evidence of forgetting over a one week delay, although floor effects were successfully avoided in most participants. This time, however, a marginally significant interaction was observed, although in the opposite direction to that found in earlier studies using integrated material (Baddeley et al., 2019; Stamate et al., 2020). Could this perhaps

reflect the unexpected occurrence of a retrieval suppression effect? This seems unlikely in view of the absence of any trace of such effects in the other four experiments using a similar design but should be noted and borne in mind should such an apparently anomalous result be found in future studies.

General Discussion

The present study forms part of an attempt to measure the rate of forgetting in individuals and hence allow the detection of ALF shown by a subsample of TLE patients. This condition is debilitating, but difficult to detect using existing standardised tests which typically only assess memory over delays of approximately 40 minutes (Baker & Zeman, 2017; Elliott et al., 2014). In order to detect ALF, patients should be tested across various delays, as its onset time and rapidity may considerably between across individuals. However, a crucial problem is that the process of testing a remembered episode is likely to actively change it, potentially either delaying forgetting as in the retrieval practice effect (Tulving, 1967; Karpicke & Roediger, 2008; Roediger & Butler, 2011; Roediger & Karpicke, 2006) or increasing forgetting when retrieval of one item actively inhibits the retrieval of others (Anderson, 2003; Anderson et al., 1994). Most of the literature on long-term forgetting avoids this problem by testing a different group of participants at each delay point, an approach that is clearly not possible when testing rate of forgetting in a single patient. Our previous attempt to solve this problem involved presenting a series of episodes and testing a separate sample of features at each delay (Baddeley et al., 2019; Stamate et al., 2020). This reduced the rate of forgetting, even though different features were sampled on each test suggesting that testing one feature of an episode had activated the remaining features.

The purpose of the present study was to test the hypothesis that retrieving one feature of an integrated episode or scene will activate the other features from that episode. This will help to protect or reactivate the memory trace, resulting in reduced forgetting on later tests. If so, we predicted that when each test item assessed a single episode (e.g. one crime) or comprised a set of isolated items (e.g. distinct words or scenes), associative links between test items would not occur and hence interpolated testing would no longer influence forgetting rate. We tested this across five experiments.

Our first experiment used a variant of the materials forming the Crimes Test for which intervening testing greatly reduced forgetting rate (Baddeley et al., 2019). By testing a separate crime at each delay (Experiment 1), we were able to minimise the between-test priming of one feature by other crimes tested at other delays. As predicted by our hypothesis, this led to equivalent rates of forgetting, regardless of intervening tests. Practical limitations in using the separate episode method led us to our next two studies using lists of individual words with a different sample of words being tested at each delay. Experiment 2 assessed memory either immediately, after 24 hours, 7 days and 28 days (multi-test group), or immediately and after 28 days (single-test group). There was no significant interaction, indicating that the rate of forgetting across the one-month delay did not significantly differ between groups. Our third study was prompted by the possibility that our null result might have resulted from floor effects at the longest delay. We therefore replicated this study but moved the final test to 7 days after encoding as opposed to 28 days. The results were essentially the same with no effect of interpolated testing on recognition memory after

a week. Our final two experiments followed the same design except that visual scenes replaced the words, again testing after a 28 day (Experiment 4) or 7 day delay (Experiment 5). The same broad pattern of results was found overall, although there was a marginally significant effect in the opposite direction to that predicted by the inter-item priming effect in Experiment 5, with interpolated testing resulting in slightly more forgetting. This is exactly opposite to the marked effects found in our earlier studies using integrated door scenes (Baddeley et al., 2019), and inconsistent with the absence of any difference in rate of forgetting shown in Experiments 1-4. However, when considering our primary research question as to whether interpolating tests *reduce* forgetting, the results of all five experiments are in line, showing that when a different episode is tested at each delay (Experiment 1) or isolated items are presented (Experiments 2-5), interpolating tests do not reduce forgetting. We next discuss the theoretical implications of these results before examining the practical implications of these and our earlier results based on integrated materials. We conclude by outlining ways in which the relative strengths of the two methods might be combined within a clinical context.

It now seems clear, across a number of experiments that when participants attempt to retrieve the whole of the material learned, retrieval will typically maintain performance, either by enhancing learning and/or reducing forgetting (Tulving 1967; Roediger & Butler, 2011; Roediger & Karpicke, 2006), the so called retrieval practice effect. Conversely, retrieval of one item may inhibit the recall of other non-recalled items (Anderson, 2003; Anderson et al., 1994). How do our own results fit in with these two superficially opposite patterns? In the case of the cued recall of integrated information such as crimes, fables and door scenes, it appears to be the case that the

cueing of one feature may activate other associated features within that episode (Baddeley et al., 2019; Stamate et al., 2020). Evidence that clusters of mutually interrelated features are forgotten less rapidly than less integrated features has been shown by others (Joensen et al., 2020; Horner & Burgess, 2013). In the case of a well-structured story or an integrated visual scene, retrieving one feature after a delay will activate and preserve the associated features, a process similar to that involved in the standard retrieval practice effect of Roediger and Karpicke (2006). The negative effects shown in retrieval induced forgetting (Anderson, 2003; Anderson et al., 1994) also stem from associations between the items, except that this time the materials are created so as to ensure competing associations to a common retrieval cue rather than mutually supportive components of an integrated episode, leading to mutual inhibition rather than facilitation. The current experiments add to this, by suggesting that when associative links between the materials tested over successive delays are avoided, neither facilitation nor inhibition is generally found.

This then raises the question of the mechanism that underpins the associative boosts observed when integrated materials are used (Baddeley et al., 2019; Stamate et al., 2020). Two clear possibilities arise, one that the process acts as a further learning trial, strengthening the memory trace and/or potentially setting up new but supportive traces. The other is the possibility that rather than depending on new learning, the effect results from reactivating the existing trace, an effect referred to as priming (Graf & Mandler, 1984; Schacter, 1992). Priming occurs when processing an item has a direct influence on subsequent behavior, either with or without the participants' conscious awareness or recollection of the priming process. This is crucially different from learning since priming tends to be preserved across a wide range of patients with

marked episodic memory deficits, where learning is impaired (Brooks & Baddeley, 1976; Cohen & Squire, 1980; Graf et al., 1984; Warrington & Weiskrantz, 1968). This distinction between episodic memory and priming is crucial for the potential usefulness of tests such as the Crimes Test and the Four Doors Test that use the sampled probing of integrated material to detect ALF. If the advantage stems from re-learning then amnesic patients characterised by a learning deficit should show a reduced benefit from interpolated testing, suggesting ALF, despite evidence that their forgetting rate is normal (Cohen & Squire,1980; Greene et al 1996; Huppert & Piercy, 1978) On the other hand, if the effect is due to priming, a widely preserved capacity across patient groups, then such patient groups should gain the same boost to retention from interpolated testing as healthy control groups. While the range of conditions tested using sampled integrated material is currently limited, the evidence points to a reduced rate of forgetting in both healthy older people and patients with Alzheimer's Disease (Stamate et al., 2020). This contrasts with evidence that patients with TLE show clear evidence of ALF on the Crimes and Four Doors Tests (Drane, 2012; Laverick, 2018). This suggests that despite the added complexity of a method reflecting two processes, both forgetting and priming, tests based on the sampling of integrated materials across delays hold considerable promise for detecting ALF.

In that case, is there any clinical need for isolated item tests of the type we have investigated? We suggest that there is, with the two approaches being complementary. As the data from individuals in both this and our earlier study (Baddeley et al., 2019) indicate, variability can be substantial, even within healthy young participant samples. This is because forgetting depends on the difference between two or more points and as such, its variance is greater than either of the two. Furthermore, given that patients

in general and TLE patients in particular may well show a difference between visual and verbal memory, potentially with only one impaired, it is clearly desirable to ensure reliability by using two measures of each modality.

At this point it may be useful to compare the two types of tests, those involving integrated material and those based on isolated items, against the criteria that Elliott et al. (2014) propose as desirable for measures of ALF. They include six that are relevant to test design, namely:

(1) There should be sufficient delay between the learning phase and the initial retention test to avoid enhanced initial memory performance based on recency. In Experiment 1, this was provided by the fourth untested crime presentation. In the other experiments in this series and Baddeley et al. (2019) using integrated materials, this was provided by interposing a recency-disrupting task between the encoding and the immediate recall test

(2) Both verbal and visual memory should be tested. This is the case for both the Crimes and Four Doors integrated cued recall tests (Baddeley et al., 2019) and for the current single item recognition tests (Experiments 2-5).

(3) Both recall and recognition should be tested. Our combined tests are one of very few that test both. Of 24 studies reported by Cassel and Kopelman (2019), only three tested recognition.

(4) Floor and ceiling effects should be avoided. This presents a potential problem for our recognition tests after the longest delay of 28 days. Given that some healthy young participants are performing at a near chance level, the test is unlikely to be informative at such long delays, although it should be suitable for delays of up to a week. In contrast, the fact that tests using integrated material greatly reduce rates of

forgetting in healthy participants and in patients whose episodic memory impairment is accompanied by preserved priming, these tests should reduce floor effects in control groups, potentially making ALF much easier to detect.

(5) Rehearsal should be prevented where possible. This is a potential problem for all methods of testing, as it is likely to be necessary to tell patients that they will be contacted and re-tested later. There may however be indirect ways of reducing emphasis on future recall and resultant possible rehearsal by including other relevant material that de-emphasises memory testing.

(6) Equate level of initial learning between patient and control groups where possible. In all four tests we aimed to use material that could be acquired relatively easily, typically in a single trial for most healthy young participants, but have also incorporated the procedure for additional trials where necessary.

All four of our tests therefore follow most of the principles proposed by Elliott et al. (2014), although they do each have limitations. The Crimes and Four Doors cued recall tests provide a less pure measure but have the advantage of minimising the degree of forgetting in both healthy people and in patients with episodic learning deficits. Our two recognition tests provide a purer measure of forgetting but show more rapid decline in healthy people which may make it more difficult to detect deviations from the norm in patients with ALF. Ideally, both should be used.

In conclusion, we began with the problem of how to detect deviations from normal rates of forgetting, which involves testing the same individual on several occasions. This in turn requires a method of avoiding interaction between the successive tests. We achieved this by using a recognition test of individual words or scenes. We

discuss both the theoretical implications of our results and their clinical applicability alongside our previous tests based on cued recall. The different methods have complementary strengths and weaknesses, suggesting it would be desirable to use all four tests to assess ALF. The material for all four of these tests are freely available on the Open Science Framework (https://osf.io/649af/?view_only=3d55c01f964d46eeb6882e9829b8e73b).

Acknowledgements

The authors would like to thank Deniz Erdil for assistance with the task development for Experiments 4 and 5, and all of the testers for their assistance with data collection.

References

Allen, R., Kemp, S., Morson, S., Wells, C., Grindheim, K., & Baddeley, A. (2019). Does telephone testing of long-term memory retention and forgetting influence performance in young and older adults? An examination using the Crimes Test. *The Neuropsychologist, 8,* 17-23.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*, 415-445.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning , Memory and Cognition, 20*, 1063-1087.

Baddeley, A., Atkinson, A., Kemp, S., & Allen, R. (2019). The problem of detecting long-term forgetting: Evidence from the Crimes Test and the Four Doors Test. *Cortex, 110*, 69-79. doi: 10.1016/j.cortex.2018.01.017

Baddeley, A., Rawlings, B., & Hayes, A. (2014). Constrained prose recall and the assessment of long-term forgetting: The case of aging and the Crimes Test. *Memory, 22*, 1052-1059. doi: 10.1080/09658211.2013.865753

Baddeley, A.D. & Hitch, G.J. (2017). Is the levels of processing effect language-limited? *Journal of Memory & Language, 92,* 1-13. doi: 10.1016/j.jml.2016.05.001

Baddeley, A. D., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the

problem of error elimination. *Neuropsychologia, 32*, 53-68.

Baker, J., & Zeman, A. (2017). Accelerated long-term forgetting in epilepsy—and beyond. In *Cognitive neuroscience of memory consolidation* (pp. 401-417). Springer.

Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology, 58*, 97-105. doi: 10.1037/h0047507

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 80*, 1-46. doi: 10.1037/h0027577

Brooks, D. N., & Baddeley, A. D. (1976). What can amnesic patients learn? *Neuropsychologia, 14*, 111-122.

Cassel, A., & Kopelman, M. D. (2019). Have we forgotten about forgetting? A critical review of 'accelerated long-term forgetting' in temporal lobe epilepsy. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 110*, 141-149. doi: 10.1016/j.cortex.2017.12.012

Cassel, A., Morris, R., Koutroumanidis, M., & Kopelman, M. (2016). Forgetting in temporal lobe epilepsy: When does it become accelerated? *Cortex, 78*, 70-84. doi: 10.1016/j.cortex.2016.02.005

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-

analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science, 210*, 207-210.

De Renzi, E., & Lucchelli, F. (1993). Dense retrograde amnesia, intact learning capability and abnormal forgetting rate: A consolidation deficit? *Cortex, 29*, 449-466.

Drane, E. (2012). Accelerated long-term forgetting off verbal material in adults with late onset temporal lobe epilepsy. Unpublished Doctoral Thesis, University of Oxford.

Elliott, G., Isaac, C. L., & Muhlert, N. (2014). Measuring forgetting: a critical review of accelerated long-term forgetting studies. *Cortex, 54*, 16-32. doi: 10.1016/j.cortex.2014.02.001

Evans, K., & Baddeley, A. D. (2018). Intention, Attention and Long-term Memory for Visual Scenes: It all depends on the scenes *Cognition, 180*, 24-37. doi: 10.1016/j.cognition.2018.06.022

Graf, P., Squire, L. R., & Mandler, G. (1984). The information that amnesic patients do not forget. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 164-178.

Green, R. E. A., & Kopelman, M. D. (2002). Contribution of recollection and familiarity judgements to rate of forgetting in organic amnesia. *Cortex, 38*, 161-178.

Greene, J. D. W., Baddeley, A. D., & Hodges, J. R. (1996). Analysis of the episodic memory deficit in early Alzheimer's Disease: Evidence from the Doors and People Test. *Neuropsychologia, 34*, 537-551.

Horner, A. J., & Burgess, N. (2013). The associative structure of memory for multi-element events. *Journal of Experimental Psychology: General, 142*, 1370-1383. doi: 10.1037/a0033626

Huppert, F. A., & Piercy, M. (1977). Recognition memory in amnesic patients: a defect of acquisition? *Neuropsychologia, 15*, 643-652. doi: 10.1016/0028-3932(77)90069-0

Huppert, F. A., & Piercy, M. (1978). Dissociation between learning and remembering in organic amnesia. *Nature, 275*, 317-318.

Isaac C.L., & Mayes, A. R. (1999). Rate of forgetting in amnesia: I. Recall and recognition of prose. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 942–962.

Jansari, A. S., Davis, K., McGibbon, T., Firminger, S., & Karpur, N. (2010). When "long-term memory" no longer means "forever": analysis of accelerated long-term forgetting in a patient with temporal lobe epilepsy. *Neuropsychologia, 48*, 1707-1715. doi: 10.1016/j.neuropsychologia.2010.02.018

Joensen, B. H., Gaskell, M. G., & Horner, A. J. (2020). United we fall: All-or-none forgetting of complex episodic events. *Journal of Experimental Psychology: General, 149*, 230-248. doi: 10.1037/xge0000648

Kapur, N., Millar, J., Colbourn, C., Abbott, P., Kennedy, P., & Docherty, T. (1997). Very long-term amnesia in association with temporal lobe epilepsy: evidence for multiple-stage

consolidation processes. *Brain and Cognition, 35*, 58-70.

Karpicke, J. D., & Roediger III, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966-968.

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology, 46*, 81-100.

Kessels, R. P. C., & de Haan, E. H. F. (2003). Implicit learning in memory rehabilitation: A meta-analysis on errorless learning and vanishing cues methods. *Journal of Clinical and Experimental Neuropsychology, 25*, 805-814.

Kopelman, M. D. (1985). Rates of forgetting in Alzheimer-type dementia and Korsakoff's syndrome. *Neuropsychologia, 23*, 623-638.

Kopelman, M. D., & Stanhope, N. (1997). Rates of forgetting in organic amnesia following temporal lobe, diencephalic, or frontal lobe lesions. *Neuropsychology, 11*, 343–356.

Laverick, T. (2018). *The use of novel measures to detect accelerated long-term forgetting: The Crimes Test and the Four Doors Test*, Unpublished doctoral dissertation, University of Durham.

Macrae, C. N., & MacLeod, M. D. (1999). On recollections lost: When practice makes imperfect. *Journal of Personality and Social Psychology, 77*, 463-473. doi: 10.1037/0022-3514.77.3.463

Mandler, G., Pearlstone, Z., & Koopmans, H. S. (1969). Effect of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 8*, 410-423.

Martin, R. C., Loring, D. W., Meador, K. J., Lee, G. P., Thrash, N., & Arena, J. G. (1991). Impaired long-term retention despite normal verbal learning in patients with temporal lobe dysfunction. *Neuropsychology, 5*, 3-12. doi: 10.1037/0894-4105.5.1.3

Mayes, A. R., Hunkin, N. M., Isaac, C. L., & Muhlert, N. (2019). Are there distinct forms of accelerated forgetting and, if so, why? . *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 110*, 115-126. doi: 10.1016/j.cortex.2018.04.005

McKee, R. D., & Squire, L. R. (1992). Equivalent forgetting rates in long-term memory for diencephalic and medial temporal lobe amnesia. *Journal of Neuroscience, 12*, 765-3772. doi: 10.1523/JNEUROSCI

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7*, 217-227.

Ricci, M., Wong, T., Nikpour, A., & Miller, L. (2019). Testing the effectiveness of cognitive interventions in alleviating accelerated long term forgetting (ALF). *Cortex, 110*, 37-46. doi: 10.1016/j.cortex.2017.10.007

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term

retention. *Trends in Cognitive Science, 15*, 20-27. doi: 10.1016/j.tics.2010.09.003

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

Schacter, D. L. (1992). Priming and multiple memory systems: Perceptual mechanisms of implicit memory. *Journal of Cognitive Neuroscience, 4*, 244-256.

Stamate, A., Logie, R. H., Baddeley, A. D., & Della Sala, S. (Epub 2020). Forgetting in Alzheimer's disease: Is it fast? Is it affected by repeated retrieval? . *Neuropsychologia, 138*, 107351. doi: 10.1016/j.neuropsychologia.2020.107351

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175-184.

Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8*, 336-342.

Warrington, E. (1984). Recognition memory test. Windsor UK: NFER-Nelson.

Warrington, E. K., & Weiskrantz, L. (1970). Amnesic syndrome: Consolidation or retrieval? *Nature, 228*, 628-630. doi: 10.1038/228628a0

Warrington, E. K., & Weiskrantz, L. (1968). New method of testing long-term retention with special reference to amnesic patients. *Nature, 217*, 972-974.