
Article:
Baddeley, AD, Atkinson, A orcid.org/0000-0001-9536-6950, Kemp, S et al. (1 more author) (2019) The problem of detecting long-term forgetting: Evidence from the Crimes Test and the Four Doors Test. Cortex, 110. pp. 69-79. ISSN 0010-9452

https://doi.org/10.1016/j.cortex.2018.01.017

Running head: Detecting long-term forgetting

The problem of detecting long-term forgetting: Evidence from the Crimes Test and the Four Doors Test

Alan Baddeley[1]

Richard Allen[2]

Amy Atkinson[2]

Steven Kemp[3]

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

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

[3] Neuropsychology, Leeds Teaching Hospitals NHS Trust, Leeds, LS1 3EX, UK

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

Abstract

While most individuals who have problems acquiring new information forget at a normal rate, there have been reports of patients who show much more rapid forgetting, particularly comprising a subsample of patients with temporal lobe epilepsy. Currently available tests are generally not designed to test this since it requires multiple different tests of the same material. We describe two tests that aim to fill this gap, one verbal, the Crimes Test, the other visual, the Four Doors Test. Each test involves four scenes comprising five features. In each case, this allows four tests of 20 different questions to be produced and used at four different delays. Two experiments were run, each comprising a multi-test condition in which immediate testing was followed by retesting after 24 hours, one week and one month, and a second condition involving a single test after one month. Both the visual and verbal tests showed clear evidence of forgetting in the single test condition, together with little evidence of forgetting in the multi-test conditions. We suggest that the testing of individual features encourages participants to remember the whole episode which then acts as a further reminder. Further research is needed to decide whether this serendipitous lack of forgetting in healthy individuals (decelerated long-term forgetting) will provide an ideal test of accelerated long-term forgetting by avoiding the danger of floor effects, or whether it will simply prove to be a further complication. Theoretical implications are discussed, as well as possible ways ahead in further investigating the surprisingly neglected field of long-term forgetting.

Keywords: Long-term forgetting, accelerated forgetting, rehearsal-induced learning, retrieval inhibition, temporal lobe epilepsy.

The work to be described resulted from an invitation from one of us (SK) to AB to join an informal group comprising neuropsychologists, neurologists and neuropsychiatrists all of whom were interested in developing measures of accelerated long-term forgetting (ALF). Although there are many neuropsychological tests concerned with long-term memory (LTM), virtually all limit long-term testing to a single re-test. This was reasonable, given that the rate of loss of information from memory appeared to be surprisingly uniform across a range of conditions including Alcoholic Korsakoff Syndrome and Alzheimer's Disease (Kopelman, 1985; Greene, Baddeley & Hodges, 1996; Huppert & Piercy,1978).

The establishment of the existence of ALF, in which apparently normal performance over short delays can be followed by a dramatic loss (Butler, Muhlert & Zeman, 2010; Butler & Zeman, 2008; De Renzi & Lucchelli, 1993; Manes, Serrano, Calcagno, Cardozo, & Hodges,2008), indicated a need for reliable measures of long-term forgetting. In an extensive review of the ALF literature, focusing particularly on patients with temporal lobe epilepsy (TLE), Elliott, Isaac and Muhlert (2014) discuss the methodological problems that confront this area, stressing the need for improved tests and procedures including parallel visual and verbal tests. A particular problem is the need to test patients repeatedly over varying delays. One option is simply to repeat the material tested originally across different delays. However, this raises two problems, the first of which is that of reliability. Tests that are designed and standardised over short delays may be highly unreliable when tested after days or weeks. Alber (2015), for example, found huge variability in the performance of healthy young participants on the Rey Auditory Verbal Learning Test (RAVLT) when retested after delays ranging from 20 minutes to one month, with some participants virtually at floor after a week while others were still performing at a high level. This

is not a criticism of the RAVLT when used in the standard way, but rather a warning that tests developed for brief delays may not be suitable for repeated testing over a longer term. A second problem concerns the effect of such successive tests on the original memory trace. Although we have been studying long-term forgetting since Ebbinghaus (1885), existing evidence provides little clear guidance on this. Long-term forgetting is a topic that has been comparatively neglected as the study of human memory moved from the forgetting-based verbal learning tradition to a cognitive tradition based on studying the processes of encoding and retrieval. Hence there is no clear agreement on when retrieval enhances subsequent recall (e.g. Roediger & Karpicke, 2006), and when it inhibits later performance (e.g. Slamecka, 1961). Some reasons for the relative lack of development of our theoretical understanding of long-term forgetting are discussed later, following the description of a more pragmatic approach to this question developed as part of the informal ALF group's communal exploration of possible ways ahead.

What follows is a brief account of the desirable characteristics of an appropriate test followed by an account of the development of two tests, one verbal and the other visual, both of which are used to investigate the influence on memory of repeated retesting. A brief consideration of the strengths and weaknesses of the two tests follows, together with suggestions as to future clinical and theoretical developments.

Desirable characteristics

So what are the desirable characteristics of a test of long-term forgetting?

- First, it should have enough capacity to provide a reliable measure across repeated tests over several separate test sessions.
- It should do so without demanding excessive initial learning time.
- The material should be patient-friendly.

- It should be able to measure performance across levels ranging from good normal memory to seriously impaired.
- The potential influence of different strategies should be minimised.
- Ideally, memory should be testable without requiring the patient to revisit the clinic, either by telephone testing or computer-based tests.

One approach to repeated testing is that taken by Cassell, Morris, Koutroumanis and Kopelman (2016) who studied verbal and visuo-spatial memory over delays between 30 seconds and a week in temporal lobe epilepsy patients. They did so by initially requiring the learning of four separate stories and four routes, then testing retention of one story and one route per delay. This has the advantage of testing each item once but has the drawback of a relatively heavy initial learning load even though a modest learning criterion of only six out of a possible ten correct answers is required. This does, however, limit potential sensitivity to scores between zero and six at each test point in some participants. A further problem is that of serial order effects during initial learning potentially favouring primacy, recency or both. This may be further complicated by test order and possible between-test interference effects. It may be the case that these effects are not problematic but further exploration with healthy participants will be needed to investigate this. In a similar vein, Jansari, Davis, McGibbon, Firminger and Kapur (2010) tested a single patient with TLE using ten stories, testing two at each of five delays, one by recall and one by recognition. From this, evidence of ALF was observed that was not found when the same story was tested repeatedly. While the study provides valuable information, the requirement to learn ten stories makes this test impracticable as a clinical measure.

A second somewhat different solution to these potential problems is offered by the cued recall method whereby retention is tested by probing the association between

features comprising the episode, as in the case of the Wrecks Test devised by Baddeley, Cuccaro, Egstrom, Weltman and Willis (1975). This was originally devised to test divers and comprised a series of verbal descriptions of scenes comprising a standard set of features: a type of ship, its name, its depth and the surrounding seabed features (for example, The fishing boat Lucky Lucy sank in 30 feet of water on a sandy bottom surrounded by kelp). Memory can then be tested by probed recall of the association between individual features. This has the advantage that the retrieval strategy is tightly constrained allowing a series of separate questions, each one of which tests one association between two features of the episode. Since TLE patients could not be assumed to have an interest in shipwrecks, our test substituted crime which television schedules suggest is of wider general interest. The resulting Crimes Test involves four incidents, each comprising a victim (age and sex), their nationality, a criminal, a crime and a location, for example, A young Chinese woman had her handbag snatched by a young girl begging outside the cathedral. Given that each feature can be probed in either direction, for example, Where was the handbag snatched? and What crime was committed outside the cathedral?, the set of four incidents generates a total of 80 different questions. These can then be split into four sets of 20 questions which can be used after varying delays. Each test set comprised an equal number of questions from each of the four crimes in a random order. Given that each association between features could be probed in either direction, we ensured that only one of these occurred within the same test. We explicitly avoided testing the four crimes separately, because any differences in difficulty between crimes would be confounded with delay, reducing sensitivity. During piloting, we also found that when tests of a single crime were grouped, participants would remember earlier responses and use these to help guess the remaining questions, a strategy that was less practical

when questions from all four crimes were randomly mixed and distributed across delays (for further details see Appendix A).

Pilot testing suggested that a single auditory presentation was sufficient to yield a high level of performance in student participants with the potential, if immediate recall is poor, of a second or third presentation in order to reach an acceptable criterion. An initial study by Baddeley, Rawlings and Hayes (2014) compared performance of young and older participants over delays ranging up to six weeks using face-to-face testing for the immediate test and telephone testing for the later tests. The results showed main effects of age and delay, together with an interaction suggesting somewhat faster forgetting in the older participants. This was unexpected, although careful searching of the literature suggested a mixed pattern of results with some suggesting equivalent rates and others more rapid forgetting in the elderly (e.g. Mary, Schreiner & Peigneux, 2013). Importantly, the principal source of the interaction was not between face to face and telephone testing but between the two delayed tests, both of which were tested by telephone. Moreover, further currently unpublished research has directly compared face to face with telephone testing with young and older participants, finding equivalent levels of performance for the two modes of test together with a small age effect, but no interaction between age and forgetting. While no significant forgetting occurred, the longest delay was a week; it is notable that the Baddeley et al. (2014) did not find a significant interaction over this timescale. Several other members of the informal ALF group have also explored the use of the Crimes Test, often finding little forgetting in healthy participants, leading to the principal empirical focus of the present study, namely the role of repeated testing. Could it be that each test functions as a rehearsal, hence maintaining memory at an initial level?

Evidence on the effect of repeated testing within the broader remit of memory studies is mixed. A number of studies suggest that retrieval of a memory trace makes other traces harder to retrieve as in the case of the part-list cueing phenomenon, whereby encouraging and practicing retrieval of one item or set of items within a list will enhance probability of its subsequent retrieval while impairing performance on non-retrieved items (Anderson 2003; Conroy & Salmon, 2006; Mueller & Brown, 1977; Slamecka, 1968). In contrast to this inhibitory effect, Tulving (1967) found that requiring participants to retrieve the whole of a previously presented list was virtually as effective as giving a subsequent learning trial. A single case autobiographical study by Linton (1975) showed that repeated recall led to better retention, while extensively replicated studies by Roediger and Karpicke (2006; Karpicke & Roediger, 2008; see Kornell & Vaughan, 2016, for a review) have shown considerable advantage on long-term retention from repeated retrievals. This effect clearly presents a potential problem for studies of ALF which in the case studied by Jansari et al. (2010) was found to mask the evidence for ALF. We ourselves hoped to avoid this whole list retrieval practice effect by ensuring that no probe question is ever repeated. The experiments that follow test this assumption.

In theoretically oriented studies, the question of whether repeated testing influences the memory trace has been finessed by testing each participant only once. This of course requires a separate group at each delay. A good example of this is provided by the elegant research program on long-term forgetting reviewed by Bahrick, Hall and Baker (2013) who tested alumni returning to the Ohio Wesleyan University annual reunions, comparing the retention, for example, of Spanish over delays ranging across many years but testing each participant only once. This design however is not possible for studies of ALF for which it is necessary to test the same

patient on several occasions, bearing in mind the fact that ALF may occur at any time and may be gradual, or relatively catastrophic. Experiment 1 studies the effect of testing on recall after one month. Healthy young participants are tested on separate parallel versions of the Crimes Test immediately and are then retested after one day, one week and one month (the multi-test condition), or are retested only after one month (the single test condition). If testing serves as further rehearsal, one would expect better delayed performance in the first group, whereas if it acts as a disruptor, the second group should be superior.

In both experiments, data were analysed using frequentist statistics as well as Bayesian factor analysis. Bayesian factor analysis compares the alternative hypothesis against the null hypothesis, and allows a test of equivalence between conditions and/or groups (Barchard, 2015; Mulder & Wagenmaker, 2016). Bayes Factors (BF) below 1 indicate evidence for the null hypothesis, whilst BFs above 1 provide evidence for the alternative hypothesis. Within ANOVA models, BFs for main effects and interactions are calculated by dividing a model including the component of interest by a model excluding the component of interest.

Experiment 1

Method

Participants. A total of 32 healthy young (age below 30) participants (22 female) were assigned at random to either a multi-test condition or a single-test condition.

Materials. Four relatively modest crimes are to be remembered and tested later. The four crimes each involve five distinctive features namely the criminal, the crime, the victim (age and sex), their nationality and the location with80questions used to

assess memory, split into four sets of 20 questions (see Appendix A). Questions were generated from the features of the crime, for example "*What crime was committed near the bridge?*", to which the answer would be "*Hit and run*". The test could also be probed in the opposite direction, for example "*Where did the hit and run occur?*", the answer being "*The bridge*". Such reversed questions never occurred in the same test. Each test comprised a mix of questions about all four crimes. The tests were presented in a counterbalanced order.

Design and procedure. A between-subjects design was employed. Participants in the multi-test condition completed an immediate test, followed by further tests involving different sets of questions after delays of one day, one week and one month. Participants in the single test condition were tested immediately and then after one month.

The test is presented as a task confronting a reporter in a small seaside town popular with tourists. The four crimes were described in turn in a series of sentences each of which was read out slowly and clearly, with a 2s pause between each sentence and a 5s pause between each crime. This was followed by a one minute interpolated task involving finding as many words as possible from the word "hippopotamus". This served to minimise any short-term recency effects.

Participants then completed an initial immediate test, which comprised one of the four sets of 20 questions. If participants scored less that 75%, the list of crimes was presented and tested again, a process that was repeated until the 75% criterion was reached. A total of 14 of the 16 participants in the repeated test group required only one trial, one required two and one required more than two. For the single test condition, 13 required one test, 3 required two and 1 more than two. The ease of learning is encouraging but the two examples of several learning trials (specific

number not recorded, both from the same tester) suggest that that the test is not yet suitable for routine clinical use without more extensive training. The cued recall test was self-paced. Participants in the multi-test condition experienced all four versions of the test, whilst participants in the single-test condition only completed tests one and four. The initial test was conducted face to face while all other tests were conducted by telephone.

A total of eight student research assistants each tested four participants, two from each group. The use of student research assistants is not typical, but can be justified on the grounds that in clinical use, tests will be given by a range of different testers with different degrees of skill and training. There was in fact a marginally significant effect of experimenter ($F = 2.45$, $p = .066$, $BF = 2.03$) suggesting that a little more training of testers might have been appropriate, but no significant interactions between experimenter and the other variables ($F \leq .974$, $p \geq .482$, $BF \leq 0.33$).

Results

Mean proportion correct (and standard error (SE)) on the immediate and one-month tests is displayed in Figure 1 as a function of test session and group. Performance at all test points, and participants' individual scores, are displayed in Figure 2.

A 2 (test session) x 2 (group) mixed ANOVA revealed significant main effects of test session ($F(1, 30) = 8.80$, $MSE = 7.39$, $p = .006$, $\eta^2p = .23$, $BF = 3.05$), in favour of the immediate test and of group ($F(1, 30) = 4.50$, $MSE = 27.21$, $p = .042$, $\eta^2p = .13$, $BF = 1.85$), with better performance by the multi-test group. These effects were qualified by a significant interaction between test session and group ($F(1, 30) = 13.37$,

MSE = 7.39, p = .001, η^2p = .31, *BF* = 30.43) , with no significant difference between groups in the immediate test session (t(30)= .26, p = .798, *BF* = 0.35), but a significant difference after one month (t(30)= 2.96, p = .006, *BF* = 7.58).

[Figure 1 about here]

Analysis was also conducted to explore whether mean proportion correct significantly differed across test sessions in the multi-test and single test conditions respectively. A one-way repeated measures ANOVA revealed no significant overall difference across test sessions in the multi-test condition (Greenhouse-Geisser corrected F(1.64, 24.61) = 0.35, *MSE* = 5.62, p = .663, η^2p = .02, *BF* = 0.12). A paired-samples t-test revealed a significant difference between test sessions in the single test condition (t(15) = 4.25, p = .001, *BF* = 52.18), with participants exhibiting higher accuracy in the immediate compared to the one month test session.

Figure 2 shows the performance across all test points, and includes results for individual participants. Data at this level are important, first in showing a relatively consistent pattern and secondly in giving no obvious evidence that lower levels of initial performance lead to faster forgetting, an issue that will be discussed later. Ease of learning the test material is reflected in the fact that 27 of the 32 participants learned the task in a single trial (although it should be born in mind that two participants required more than two trials and that these are young and healthy student participants).

[Figure 2 about here]

After the final test, participants were asked if they had rehearsed the material at any point. A total of four participants in the multi-test and three in the single test condition reported rehearsing at least once. The three rehearsers in the single test condition all improved as did three of the four in the multi-test group. These cases were removed and the data re-analysed. The effect of test session remained significant ($F(1, 23) = 28.38$, $MSE = 4.78$, $p < .001$, $\eta^2p = .55$, $BF = 35.35$), as did the effect of group ($F(1, 23) = 5.29$, $MSE = 28.70$, $p = .031$, $\eta^2p = .19$, $BF = 2.49$) and the interaction ($F(1, 23) = 22.15$, $MSE = 4.78$, $p <. 001$, $\eta^2p = .49$, $BF = 129.15$). Nevertheless, the possibility of rehearsal is clearly an important issue if the test is to be taken further. Discussion of this issue follows Experiment 2, which applies a similar design to a second potential test of ALF, based on the cued recall of visual stimuli.

Experiment 2

This followed the development of a visual equivalent of the Crimes Test, and comprised four door scenes, each involving five distinct features. Pilot testing showed that healthy young participants required 10s per door to reach a level of approximately 80% correct on probed recall.

Method

Participants. Forty healthy young (below age 30) participants (27 female) were randomly allocated: 20 to the multi-test and 20 to the single test condition.

Materials. Four door scenes were created, each comprising five distinct features: the type of door (house, factory, gate or church), the door colour (yellow, green, black or red), the colour of its surround (black, red, yellow or white), the object above it (a window, star, balcony or statue) and the creature in front of it (cat, pig, dog

chicken; see Figure 3). Presentation order was always: house door, factory, gate and church.

As in Experiment 1, 80 questions were created to assess memory for the doors, such as *"What colour is the church door?"* and *"What creature is in front of the door with the yellow surround?"*. These questions were randomly placed into four sets of 20 questions, avoiding within-list reversed versions of the same association. The tests were administered in a counterbalanced order.

[Figure 3 about here]

Design and Procedure. A between-subjects design was employed. As in Experiment 1, participants in the multi-test condition completed an immediate test, followed by further tests involving different questions at delays of one day, one week and one month. Participants in the single test condition were tested immediately and then after one month.

Firstly, participants completed the encoding phase, in which each door was displayed for ten seconds. Before each door was presented, participants were told the type of door, and asked to repeat the name three times (e.g. church, church, church). They then pressed a button to display the door. During presentation, participants were asked to repeat the type of door a further five times, whilst also trying to remember the other features. This procedure served two functions, making clear the type of door and discouraging verbal rehearsal of the features; a strategy that is likely to be very unhelpful.

Immediately after the door was displayed, participants were asked to name all five features. If participants recalled any features incorrectly, they were shown the

door again for a further two seconds for each feature inaccurately recalled. The door was then removed and participants were asked to name the features previously recalled incorrectly. If participants incorrectly recalled a feature twice, they were shown the door a third time, and the correct answer was highlighted by the experimenter. After all four doors had been displayed, participants were given one minute to note down as many words as possible from 'hippopotamus'.

Participants then completed an immediate test of the doors, in which they were asked one of the four sets of questions. If participants answered 75% or more of these questions correctly, the initial test session ended. If participants scored below this cut off, they were shown the doors and asked to complete the immediate test phase again. This continued until participants answered 75% or more of the questions correctly. The questions asked in these subsequent rounds of initial testing were the same as those asked in the immediate test. Of the 18 participants in the multi-test condition for whom this information was recorded,11 required only one presentation, 5 required two and one each required 3 and 4 presentations. In the single test condition, of 18 for whom this was reported, 8 required only a single presentation, 9 required two and 1 required three. Comparison across conditions using Chi-squared indicated no difference between conditions $\chi^2 = 2.617$, df £ p = 0.455. Follow-up test sessions were completed over the telephone.

Data were collected from eight student research assistants as well as two more experienced testers. There was, however, no significant effect of experimenter ($F = 1.11$, $p \geq .482$, $BF = 0.22$) and no interaction between experimenter and other variables ($F \leq .815$, $p = .395$, $BF \leq 0.38$).

Results

Mean proportion correct (and SE) at immediate testing and after one month are displayed in Figure 4 as a function of test session and group. Data from individual participants are shown in Figure 5.

[Figure 4 about here]

A 2 (group) x 2 (test session) mixed ANOVA was conducted. This indicated significant main effects of test session ($F(1, 38) = 60.17$, MSE = .02, $p < .001$, η^2p =.61, BF > 1000), indicating overall evidence of forgetting, and of group ($F(1, 38) = 16.92$, MSE = .03, $p < .001$, η^2p =.31, BF = 81.14). These effects were qualified by a significant interaction between group and test session ($F(1, 38) = 14.88$, MSE = .02, $p < .001$, η^2p =.28, BF = 77.50). Subsidiary analysis revealed no significant difference between groups at immediate testing ($t(38) = 1.75$, $p = .088$, BF = 1.03), but a significant difference at the 28-day test session ($t(38)=4.40$, $p < .001$, BF = 249.49), indicating that the forgetting effect was principally attributable to the single test group.

A one-way repeated measures ANOVA was conducted to explore whether the proportion of questions answered correctly significantly differed between the four test sessions in the multi-test condition. This revealed a significant effect of test session ($F(3, 57) = 4.53$, MSE = .01, $p = .006$, η^2p =.19, BF = 6.50). Bonferonni post-hoc comparisons revealed a significant difference between immediate testing and 28-day test session ($p = .031$), but no other significant differences ($p > .05$). A paired samples t-test was also conducted to explore whether the proportion of questions answered correctly significantly differed in the single test condition between immediate testing and the one-month test session. This revealed a significant effect of test session ($t(19)$

= 7.39, $p < .001$, BF > 1000), with participants answering significantly more questions correctly at immediate testing than after one month.

[Figure 5 about here]

General Discussion

Both Experiments 1 and 2 show the same overall pattern of relatively well-maintained performance across the one month delay in the multi-test condition in contrast to a marked loss in the single test condition. This clearly supports the idea that repeated cued recall tests help maintain performance, as found in studies where complete recall was required (Tulving, 1967; Jansari et al., 2010; Karpicke & Roediger, 2008), but in contrast to the negative effects that can be found in part-list retrieval situations in which probing recall of one item can inhibit subsequent recall of the remaining items (Anderson, 2003; Conroy & Salmon, 2006; Mueller & Brown, 1957; Slamecka, 1961). This in turn suggests that, although memory was tested using specific individual questions, with no question asked more than once, the situation behaved as though the whole episode were being tested. We suspect that the process of retrieving a specific piece of information may typically involve retrieving the episode in which it was embedded, a process that would enhance later performance on other features that were not specifically tested at that point. This interpretation has similarities to what Kornell and Vaughan (2016) describe as a search set theory whereby test trials activate not only the item cued but also associated information. In this connection, it would be interesting to run parallel studies using lists of independent items comprising words or pictures of objects in which testing a single item would be less likely to evoke retrieval of surrounding items. It seems possible that, in these

circumstances, evidence of retrieval-based refreshing might be replaced by the opposite phenomenon of retrieval-based inhibition.

This is clearly an issue of theoretical importance, but from a practical viewpoint our results could be serendipitous. If the principal aim is to measure as purely as possible the decline of memory over time, multiple testing of the same person is clearly unsatisfactory since the process of testing appears to refresh the memory, hence making the classic between-groups design in which different individuals are tested at different delays much preferable. However, if the aim is to detect ALF, then multiple tests on the same individual are likely to be necessary and from this viewpoint the fact that healthy people show little forgetting could potentially offer a clear advantage.

This in turn raises the question of the mechanism underpinning the multi-test effect; is it simply refreshing the existing representations, or does it represent new learning? If it represents new learning, then patients with impaired learning capacity but normal rates of forgetting are likely to be penalised with their learning deficit potentially mistaken for faster forgetting. In this case, classic amnesic syndrome patients would appear to show ALF. If on the other hand, it reflects an implicit priming effect which is typically preserved in such patients (Brooks & Baddeley, 1976; Schacter & Graf, 1986; Squire, 1992), they should show relatively preserved long-term implicit memory performance, as was found for example in the case of the Hebb repeated digit sequence paradigm by Baddeley & Warrington (1970). We clearly need further evidence from other patient groups who show memory deficits, to ensure that their learning impairment does not result in faster forgetting under multi-test conditions. Conversely it is also important to use the multi-test paradigm to assess

well-studied patients who have already shown demonstrable ALF, to ensure that the predicted forgetting occurs under multi-test conditions.

Existing evidence on this general issue is currently fragmentary. If the refreshing effect of retesting reflects learning capacity then one might expect poor learners to show faster forgetting. This clearly requires careful further investigation, but inspection of Figures 2 and 5 do not suggest obviously faster forgetting for poorer performers, provided ceiling effects are avoided. This conclusion was supported when the groups were split into those performing above and those below the median. The two groups showed essentially parallel forgetting functions, although with such small numbers a statistical analysis demonstrating the absence of a significant interaction would carry little weight. Rather more worrying from a clinical viewpoint is the fact that a small number of the healthy young participants on each test do show substantial forgetting, again suggesting that more development of the test is needed before use clinically.

The evidence for faster forgetting in the elderly by Baddeley et al. (2014) might favour the idea that testing involves relearning with this being more limited in the elderly This result is not, however, typical of existing literature where Salthouse (1991) reviewing 22 studies found significant evidence of faster forgetting in only half of them, while a meta-analysis by Kausler (1991) also found inconsistency. Finally, a currently unpublished D.Clin.Psych thesis by Drane (2014) using the Crimes Test showed an encouraging pattern of results with marked forgetting over time in a temporal lobe epilepsy group together with a very flat function in healthy controls. We clearly need more data on patient groups with memory deficits on the one hand, and on patients who show clear evidence of ALF on the other.

Whatever such studies reveal, there are still a number of practical problems that need to be faced in the area in general and with these two tests in particular. One concerns the level of difficulty of the initial learning task. It is important that the tests should be suitable for a wide range of patients, including those with general problems in learning and memory. Both tests were designed to be relatively easy for healthy participants and were acquired in one or two trials by most, though not all participants, but are likely to be less so for patients, introducing the problem of a possible interaction of forgetting rate and initial performance level. Our current strategy is to use additional presentations if initial performance is low. However, this leads to two questions, the first being whether even this will produce a reasonably high level of performance, while the second concerns whether rate of forgetting is indeed influenced by level of initial learning. This is clearly a general issue for the field of forgetting.

A further practical problem for these and potentially other tests of long-term forgetting concerns the issue of rehearsal. Data from Experiment 1 suggests that a minority of participants rehearsed at least once (20%), with the three individuals who were in the single-test condition all proving to be atypical in showing a slight improvement over the delay. Unfortunately, we did not question participants in the Four Doors study (which was in fact run before its equivalent involving the Crimes Test). This is clearly a potential problem in any situation where repeated testing is used. Whether rehearsal will be at all effective in preventing ALF, however, remains to be seen. We suspect that this may be less of problem when isolated items rather than episodes and scenes are used; we plan to investigate this.

In conclusion, we have developed two tests which have the unexpected characteristic of showing very little forgetting over subsequent tests in healthy

participants, in effect, decelerated long-term forgetting. Whether the absence of complications due to changes in overall performance levels over time means that the tests are particularly appropriate for measuring ALF or whether the opposite is the case, clearly requires further investigation.

Acknowledgements

We are grateful to fellow members of the informal Accelerated Long-term Forgetting group for valuable discussion and permission to quote their as yet unpublished findings, and to the undergraduate volunteers without whose help Experiments 1 and 2 would not have been possible.

Highlights

- Detecting long-term forgetting may need repeated tests of the same material.
- Current standard tests are not well designed for this.
- The verbal Crimes Test and the visual Four Doors Test yielding 4 sets of 20 questions.
- We compared multi-test and single test probe recall over a one month delay.
- Repeated testing avoided the clear forgetting shown in the single test condition.

Figure captions

Figure 1. Mean proportion correct (and SE) on the Crimes Test in the immediate and one-month tests as a function of test session and group.

Figure 2: The proportion of questions each participant answered correctly during the test sessions in each test condition. Thick black lines represent the mean (with error bars denoting SE).

Figure 3: The images of doors used. The doors were shown one by one in the same order to all participants.

Figure 4: Mean proportion correct (and SE) on the immediate and one-month Four Doors Tests as a function of test session and group.

Figure 5: The proportion of questions each participant answered correctly during the test sessions in each test condition. The thicker black lines represent the mean (with the error bars denoting SE).

Appendix A

Revised Crimes Test

Imagine you are a reporter on a newspaper serving a small coastal town, popular with tourists and suffering from a range of minor crimes. You check in on a Monday morning and your editor asks you to investigate the following four incidents. I want you to try to see them in your mind's eye as I describe them and remember the basis features of each, who did what to whom and where. You probably won't remember all of it, but do your best.

Any questions?

Then here are the crimes:
When reading, pause for 2 seconds after each sentence (.) and for 5 seconds between incidents

---

- An elderly Indian man went into a pub after a day of sightseeing.
- He hung up his jacket and ordered a beer.
- He was watching the sun going down when he noticed that a tramp was stealing his coat.

- A young Chinese woman had arranged to meet her sister at morning service.
- As she was about to enter the church she noticed a young girl who seemed to be begging.
- She suddenly snatched the woman's handbag and ran off.

- A young Frenchman was leaving a nightclub in the early hours of the morning.
- As he walked down a street near the docks, a shadowy figure approached and offered to sell him drugs.
- He refused whereupon the drug dealer stabbed him and ran off.

- An old Russian lady was walking back to her hotel across the river.
- As she approached the bridge a speeding car veered onto the pavement and hit her.
- The driver, a teenage, girl, leapt out and ran away.

Test A

1	What was the age/sex of the victim of the stabbing crime? *Young man*	
2	Who committed the crime against the Indian person? *tramp*	
3	What was the age/sex of the victim from Russia? *Old lady*	
4	What was the crime committed against the young woman? *Handbag snatch*	
5	What was the crime committed by the drug dealer? *stabbing*	
6	What was the age/sex of the victim of the crime committed by the young girl? *Young woman*	
7	What was the location of the stabbing? *docklands*	
8	Who committed the crime at the pub? *tramp*	
9	What was the crime committed on the bridge? *Hit & run*	
10	What was the nationality of the victim of the stabbing? *French*	
11	What was the age/sex of the victim of the crime at the church? *Young woman*	
12	What was the location of the crime committed by the tramp? *pub*	
13	What was the age and sex of the person who committed the crime against the old lady? *Teenage girl*	
14	What was the age and sex of the person who committed the crime against the young woman? *Young girl*	
15	What was the location of the hit and run crime? *bridge*	
16	What was the crime committed against the old man? *Coat theft*	
17	What was the location of the crime committed against the person from Russia? *bridge*	
18	What was the location of the handbag snatch? *church*	
19	Who committed the crime against the French person? *Drug dealer*	
20	What was the nationality of the victim in the pub? *Indian*	

Test B

1	What was the age/sex of the victim of the coat stealing? *Old man*	
2	What was the location of the stabbing crime? *docklands*	
3	What was the crime committed against the Russian person? *Hit and run*	
4	What was the age/sex of the victim from China? *Young woman*	
5	What was the nationality of the young man? *French*	
6	What was the age/sex of the victim of the crime committed by the teenage girl? *Old lady*	
7	What was the location of the coat theft? *pub*	
8	Who committed the crime at the docklands? *Drug dealer*	
9	What was the nationality of the victim of the crime near the church? *Chinese*	
10	What was the nationality of the victim of the coat theft? *Indian*	
11	What was the age/sex of the victim of the crime on the bridge? *Old lady*	
12	What was the location of the crime committed by the young girl? *church*	
13	Who committed the crime against the young man? *Drug dealer*	
14	What was the nationality of the young woman? *Chinese*	
15	What was the nationality of the victim of the crime committed by the tramp? *Indian*	
16	Who committed the hit and run crime? *Teenage girl*	
17	What was the location of the crime committed against the person from China? *Church*	
18	What was the crime committed against the person from France? *Stabbing*	
19	What was the crime committed against the person from India? *Coat theft*	
20	What was the nationality of the victim of the crime committed on the bridge? *Russian*	

Test C

1	Who committed the coat stealing? *Tramp*	
2	What was the location of the coat theft? *pub*	
3	Who committed the crime against the Chinese person? *Young girl*	
4	What was the crime committed against the old lady? *Hit & run*	
5	What was the nationality of the old man? *Indian*	
6	What was the age/sex of the victim of the crime committed by the drug dealer? *Young man*	
7	Who committed the stabbing? *Drug dealer*	
8	Who committed the crime at the church? *Young girl*	
9	What was the crime committed near the docklands? *stabbing*	
10	What was the nationality of the victim of the handbag snatch? *Chinese*	
11	What was age/sex of the victim of the crime at the pub? *Old man*	
12	What was the location of the crime committed by the teenage girl? *bridge*	
13	What was the crime committed by the tramp? *Coat theft*	
14	What was the nationality of the victim of the crime committed by the teenage girl? *Russian*	
15	What was the location of the hit and run? *bridge*	
16	What was the location of the crime committed against the person from France? *docklands*	
17	What was the location of the crime committed against the young woman? *church*	
18	Who committed the crime against the person from Russia? *Teenage girl*	
19	What was the nationality of the victim of the crime committed by the drug dealer? *French*	
20	What was the crime committed by the young girl? *Handbag snatch*	

Test D

1	What was the age/sex of the victim of the handbag snatching? *Young woman*	
2	What was the crime committed in the pub? *Coat theft*	
3	Who committed the handbag snatching? *Young girl*	
4	What was the crime committed against the young man? *stabbing*	
5	What was the nationality of the old lady? *Russian*	
6	What was the age/sex of the victim of the crime committed by the tramp? *Old man*	
7	What crime was committed against the person from China? *Handbag snatch*	
8	Who committed the crime on the bridge? *Teenage girl*	
9	What was the crime committed near the church? *Handbag snatch*	
10	What was the nationality of the victim of the hit and run? *Russian*	
11	What was the age/sex of the victim of the crime committed in the docklands? *Young man*	
12	What was the nationality of the victim of the crime committed by the young girl? *Chinese*	
13	Who committed the crime against the old man? *Tramp*	
14	What was the crime committed by the teenage girl? *Hit & run*	
15	What was the nationality of the victim of the crime in the docklands? *French*	
16	What was the age/sex of the victim from France? *Young man*	
17	What was the location of the crime committed against the person from India? *pub*	
18	What was the age/sex of the victim of the hit and run crime? *Old lady*	
19	What was the age/sex of the victim from India? *Old man*	
20	What was the location of the crime committed by the drug dealer? *Docklands*	

References

Alber, J. (2015). *Improving longer-term memory via wakeful rest in health and amnesia: Evidence for memory consolidation.* (unpublished PhD thesis), University of Edinburgh.

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

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., Cuccaro, W. J., Egstrom, G. H., Weltman, G., & Willis, M. A. (1975). Cognitive efficiency of divers working in cold water. *Human Factors, 17*, 446-454.

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 176-189.

Bahrick, H. P., Hall, L. K., & Baker, M. K. (2013). *Life-Span Maintenance of Knowledge* New York: Psychology Press.

Barchard, K. A. (2015). Null Hypothesis Significance Testing Does Not Show Equivalence. *Analyses of Social Issues and Public Policy*, *15*(1), 418-421. doi: 10.1111/asap.12095

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

Butler, C. R., Mulhert, N., & Zeman, A. (2010). Accelerated long-term forgetting. In S. Della Sala (Ed.), *Forgetting* (pp. 211-237). Hove: Psychology Press.

Butler, C. R., & Zeman, A. (2008). Recent insights into the impairment of memory in epilepsy: Transient epileptic amnesia, accelerated long-term forgetting and remote memory impairment. *Brain, 131*, 2243-2263.

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

Conroy, R., & Salmon, K. (2006). Talking about parts of a past experience: The influence of elaborative discussion and event structure on children's recall of nondiscussed information. *Journal of Experimental Child Psychology, 95*, 278-297.

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

Drane, E. (2014). *Accelerated long-term forgetting of verbal material in adults with late-onset temporal lobe epilepsy*. DClinPsych. Dissertation. University of Oxford. unpublished.

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. New York: Dover.

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

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.

Huppert, F. A., & Piercy, M. (1978). The role of trace strength in recency and frequency judgements by amnesic and control subjects. *Quarterly Journal of Experiment Psychology, 30*, 346-354.

Jansari, A. S., Davis, K., McGibbon, T., Firminger, S., & Kapur, 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

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

Kausler, D. H. (1991). *Experimental psychology, cognition and human aging* (2nd ed.). New York: Springer-Verlag.

Kopelman, M. D. (1985). Rates of forgetting in Alzheimer-type dementia and Korsakoff's syndrome. *Neuropsychologia, 23*, 623-628.

Kornell, N., & Vaughan, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation, 65*, 183-215. doi:http://doi.org/10.1016/bs.plm.2016.03.003

Linton, M. (1975). Memory for real-world events. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. Chapter 14). San Francisco: Freeman.

Manes, F., Serrano, C., Calcagno, M. L., Cardozo, J., & Hodges, J. (2008). Accelerated forgetting in subjects with memory complaints: A new form of mild cognitive impairment? *Journal of Neurology, 255*, 1067-1070. doi:10.1007/s00415-008-0850-6

Mary, A., Schreiner, S., & Peigneux, P. (2013). Accelerated long-term forgetting in aging and intra-sleep awakenings. *Frontiers in Psychology, 4*, 750. doi:10.3389/fpsyg.2013.00750

Mueller, J. H., & Brown, S. C. (1977). Output interference and intralist repetition in free recall. *American Journal of Psychology, 90*, 157-164.

Mulder, J., & Wagenmakers, E. J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1-5. doi: 10.1016/j.jmp.2016.01.002

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

Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging*. Hillsdale, NJ: Erlbaum.

Schacter, D. L., & Graf, P. (1986). Preserved learning in amnesic patients: Perspectives from research on direct priming. *Journal of Clinical and Experimental Neuropsychology, 8*, 727-743.

Slamecka, N. J. (1961). Proactive inhibition of connected discourse. *Journal of Experimental Psychology, 62*, 295-301.

Slamecka, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology, 76*, 504-513.

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience, 4*, 232-243.

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.

Figure 1

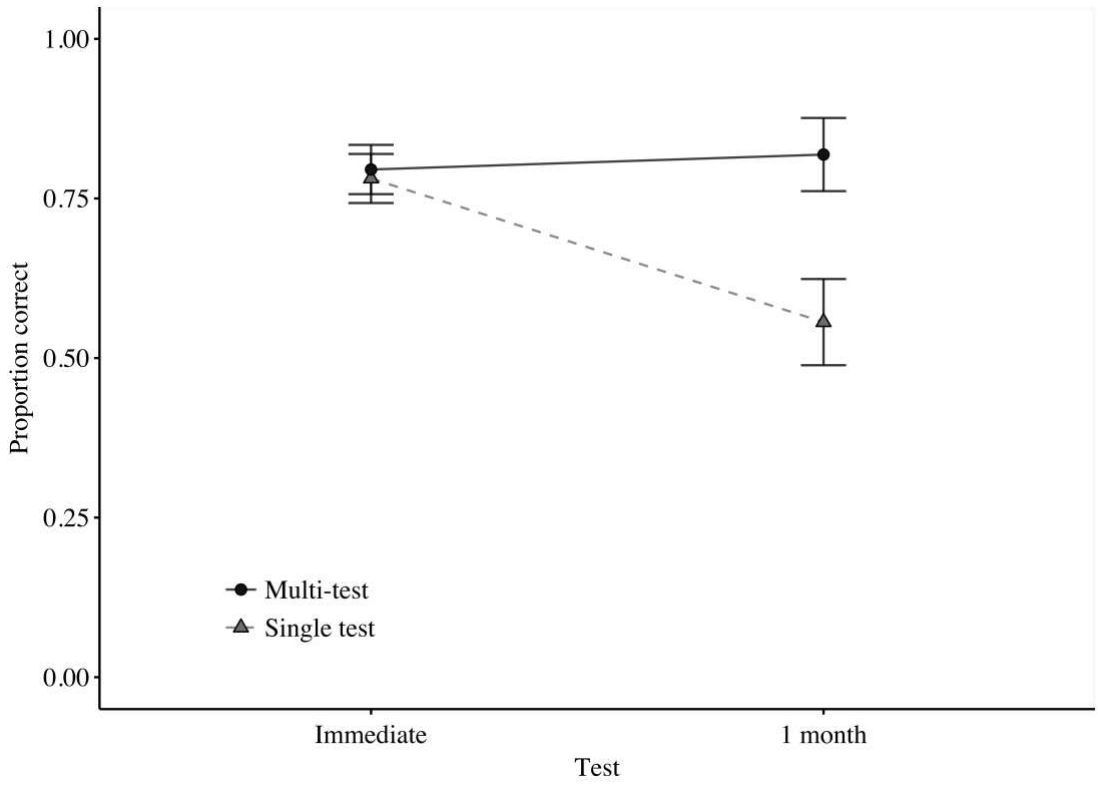

1.00
0.75
0.50
0.25
0.00
Proportion correct
Multi-test
Single test
Immediate
1 month
Test

Figure 2

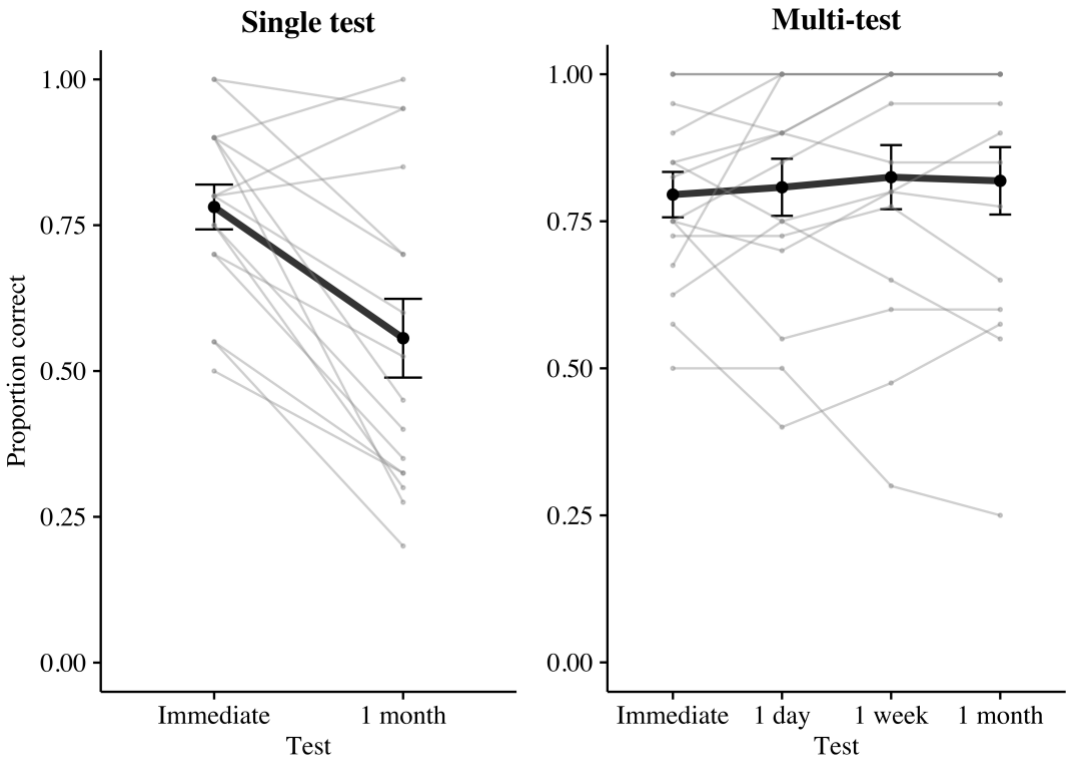
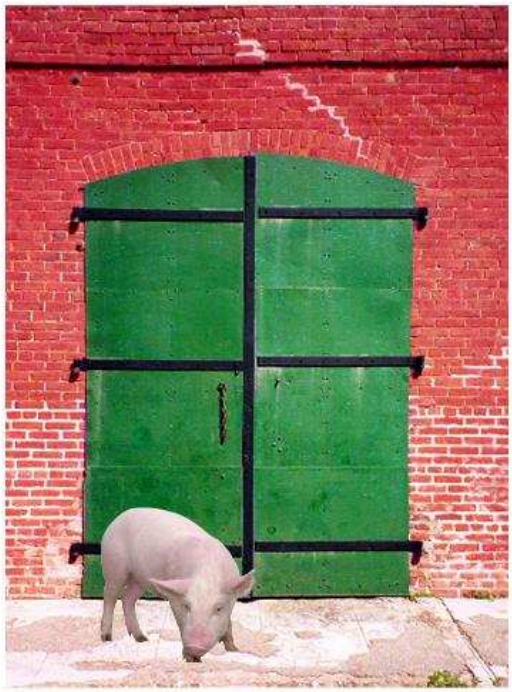

Single test
Multi-test
Proportion correct
1.00
0.75
0.50
0.25
0.00
Immediate
1 month
1 day
1 week
Test

Figure 3

HOUSE



FACTORY

GATE

CHURCH

Figure 4

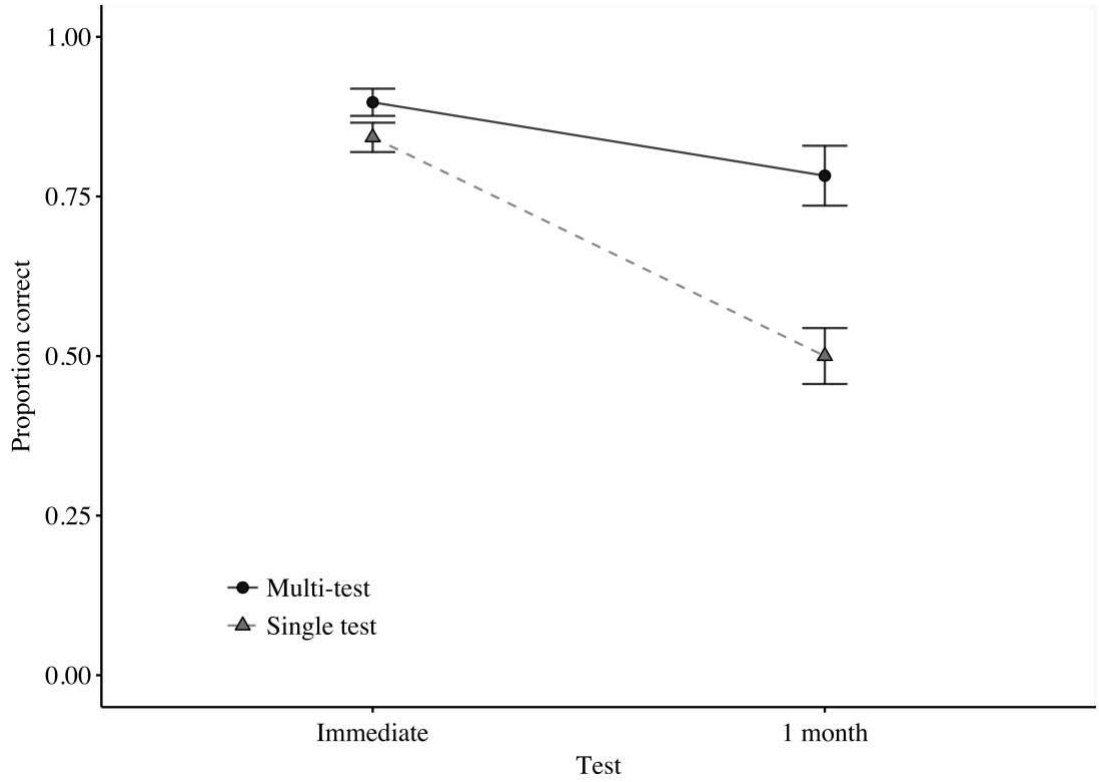

1.00
0.75
0.50
0.25
0.00
Proportion correct
Multi-test
Single test
Immediate
1 month
Test

Figure 5