# Accelerated Long-Term Forgetting: Development of a sensitive measure and its application to unanswered questions in forgetting research.

## PhD Thesis

**Student ID**: 33492064, Terence McGibbon

Department of Psychology Goldsmiths, University of London

Supervised by Dr Ashok Jansari

Target submission date: 31st January 2023

Word count: 93,824

# Author's Declaration

The data for Experiments 1 and 2 were gathered by MSc students Ana Nemes and Jessica Demirjian respectively, as part of their MSc projects. The author assisted them with construction of their experimental designs. The analyses of the data presented here has been performed by the author, correcting errors in their analyses and extending the analyses to consider questions relevant to this thesis. All content in the discussion of these experiments is the author's own.

Some of the contents of Chapter 2 have been accepted for publication prior to the submission of this thesis, as McGibbon et al. (2022).

Professor Alan Pickering at Goldsmiths, University of London has validated the analysis and guidelines developed in Chapter 4 and has expressed interest in being a co-author on a methodology paper based on this work.

# Acknowledgments

The author would like to thank his supervisor, Dr. Ashok Jansari, for his ongoing advice and guidance throughout this PhD, and without whom this work would not have been possible.

The author is grateful for the help of Professor Josh Davis at the University of Greenwich, who assisted with recruitment of participants for Experiment 6 from the University of Greenwich participant database.

The author would also like to thank the numerous participants who have taken part in this research.

# Abstract

Accelerated Long-term Forgetting (ALF) is a disorder in which new information can be learnt and retained normally over short delays, but is then forgotten at an accelerated rate at longer delays. It was first detected in epilepsy, but has since been reported in healthy older individuals, and is a possible early marker for risk of developing Alzheimer's disease (AD). Research has been hampered by lack of a methodologically sound test with adequate sensitivity which can be used in both research and clinical settings. In the current research a novel measure, the Verbal Associative Learning and Memory Test (VALMT), was developed from an initial face-to-face test into a fully automated online test, with parameters tuned to maximise sensitivity and avoid ceiling effects. In parallel, a novel set of practical guidelines were developed to assist in experimental design to ensure ceiling effects do not influence results.

VALMT was used to investigate three topics: ALF symptoms in healthy ageing and the possibility of using ALF as a marker for those at risk of developing AD; memory performance over extended delays in the general population; the timeframe of onset of accelerated forgetting. This research showed that VALMT provides a sensitive measure of accelerated forgetting that can be used for all ages from 16 to 82, and is more sensitive than a standard clinical test (Wechsler Memory Scale Logical Memory). Testing across the lifespan indicated a gradual decline in memory performance starting in the 57-69 age band. Importantly, this research also highlighted the existence of a subset of healthy older participants who learn more slowly and forget more rapidly, and indicates the potential for this learning deficit and subsequent forgetting to be used to identify older individuals who are at risk of developing dementia, which will be of significant benefit in research and clinical practice.

# Contents

# List of tables

# List of figures

# 1 Chapter 1 – Study space analysis and literature review

## 1.1 Introduction

The study of forgetting has a long history in psychology, going back at least to the studies of Ebbinghaus (1885). However, this field has seen increased interest in the last 20 years with conferences and meetings now being dedicated to the topic (e.g., EPS Forgetting Workshop, 2017). One reason for this is the identification of Long-Term Amnesia, later renamed Accelerated Long-term Forgetting (ALF), as a distinct memory complaint (e.g. Kapur et al., 1997). ALF is typically defined as a disorder in which new information can be learnt and retained normally over at least 30mins (the delay used in standard anterograde memory tests), but is then forgotten at an accelerated rate at longer delays. Initial ALF research focused on people who report being able to remember events and information for several days, but then display an accelerated forgetting which becomes noticeable after a few days to weeks (e.g. Butler et al., 2007). This is of clinical interest as it highlights a class of patient who may perform normally on standard clinical tests of anterograde memory, while still having a genuine memory disorder. It is also of theoretical interest; for example, Butler et al. argue it provides evidence for secondary consolidation processes, occurring well after initial encoding (Alvarez & Squire, 1994; Nadel & Moscovitch, 1997), which when disrupted can lead to symptoms of late-onset accelerated forgetting.

While early papers in the field had documented ALF in patients with epilepsy with a temporal lobe focus, particularly Temporal Lobe Epilepsy (TLE) or Transient Epileptic Amnesia (TEA), more recent papers have shown that it can also be found in other populations. For example, Gascoigne et al. (2012) have shown that ALF can occur with epilepsy not restricted to the temporal lobes; they have documented ALF in 20 children with Idiopathic Generalised Epilepsy (IGE). More recently, Lah et al. (2017) found ALF in children who had sustained Traumatic Brain Injury (TBI) where, importantly, one of the exclusion criteria was any history of seizures preceding or post the TBI. ALF has also been identified in healthy older adults (Baddeley et al., 2014) and in older participants who complained of memory problems but who performed normally on standard clinical tests (Manes et al., 2008). However, ALF is still a relatively new field, with many unresolved issues and possibilities for impactful research.

## 1.2 Initial study space analysis and selection of research areas

To formally assess the coverage of existing ALF research and identify opportunities for PhD level research a study space analysis (Malpass et al., 2008) was performed in November 2017, summarising ALF literature published to that date. It should be noted that this was a point in time exercise, intended to identify candidate areas for research; its inclusion here is not intended to provide an up to date summary of the research base as more recently work published during the progression of this PhD is not included. The results are summarised in Table 1.1. For any particular combination of study parameters the relevant cell in the table shows the number of published studies. For example, to find the number of studies using individuals with MCI/SCI as the population (item 3.10), and using visual figures as the stimuli (item 5.4) find the cell at row 3.10 and column 5.4, which shows one existing published study.

Table 1.1 highlights those areas where extensive research has already been performed, and several areas were research is sparse or non-existent. New research will, in general, be most impactful where it addresses under researched topics, rather than those with a solid existing research base. A large proportion of existing ALF research focuses on various forms of epilepsy. In contrast, some other key populations and factors show a lack of coverage, and are therefore prime targets for new research as part of a PhD. From these, the following were identified as primary candidate research areas for this PhD, based on the combination of a lack of existing research and the potential importance of new research findings (relevant areas in Table 1.1 are highlighted in yellow):

- the distribution of ALF in the general population and, more generally, the development of memory performance at extended delays across the entire adult lifespan
- presence of ALF in healthy ageing and in Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD), including the use of ALF as an early marker of those at risk of developing MCI/AD

*Table 1.1 Study space analysis summary*

| | | 1.1 | 1.2 | 1.3 | 2.1 | 2.2 | 2.3 | 3.1 | 3.2 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 3.10 | 3.11 | 3.12 | 3.13 | 3.14 | 4.1 | 4.2 | 4.3 | 4.4 | 4.5 | 5.1 | 5.2 | 5.3 | 5.4 | 5.5 | 5.6 | 5.7 | 6.1 | 6.2 | 6.3 | 6.4 | 6.5 | 6.6 | 6.7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.1 | case studies | | | | | 5 | 3 | 5 | 3 | | 1 | | 1 | 2 | | 4 | | | | | | 6 | 3 | | 1 | 1 | 4 | | 5 | 4 | 1 | | 1 | 6 | 1 | | 1 | 5 | 5 | 4 |
| 1.2 | group studies | | | | 1 | 16 | 18 | 13 | 6 | 6 | 10 | 8 | 1 | | | 14 | 3 | 7 | | | 7 | 28 | 18 | | | 2 | 11 | 2 | 17 | 12 | 5 | 2 | 5 | 28 | 9 | 2 | 13 | 21 | 11 | 18 |
| 1.3 | general population | | | | | 2 | 2 | | | | | | | | | | | | | | | 2 | 1 | | | | 1 | 1 | | | | | | 1 | | | 1 | 2 | 1 | 1 |
| 2.1 | children | | | | | | | | | 1 | | | | | | | | | | | | 1 | 1 | | | | | | 1 | | | | | 1 | | | | 1 | | |
| 2.2 | adults | | | | | | 4 | 17 | 2 | 5 | 11 | 9 | 2 | 1 | | 17 | | | | | 5 | 20 | 13 | | | 2 | 12 | 1 | 11 | 8 | 4 | 2 | 3 | 18 | 4 | | 8 | 15 | 8 | 10 |
| 2.3 | elderly | | | | | | | 2 | 8 | | | | | 1 | | 4 | 3 | 7 | | | 2 | 16 | 10 | | 1 | 1 | 5 | 2 | 10 | 8 | 4 | | 3 | 18 | 7 | 2 | 7 | 14 | 10 | 14 |
| 3.1 | TLE | | | | | | | | 1 | 5 | 11 | 8 | 2 | 1 | | 16 | | | | | 5 | 18 | 12 | | | 2 | 12 | | 10 | 8 | 3 | 2 | 3 | 16 | 3 | | 7 | 11 | 7 | 8 |
| 3.2 | TEA | | | | | | | | | | | 1 | | 1 | | 3 | | | | | 2 | 8 | 2 | | 1 | 1 | 1 | 1 | 7 | 4 | 1 | | 2 | 8 | 2 | 2 | 3 | 8 | 7 | 8 |
| 3.3 | ETE/IGE | | | | | | | | | | 4 | 4 | | | | 5 | | | | | 2 | 6 | 5 | | | | 3 | | 3 | 2 | 1 | | 1 | 6 | | | 3 | 6 | | 3 |
| 3.4 | seizures/subclinical | | | | | | | | | | | 6 | 2 | 1 | | 9 | | | | | 4 | 10 | 7 | | | 2 | 6 | | 4 | 3 | 2 | 1 | 2 | 9 | 2 | | 5 | 6 | 3 | 5 |
| 3.5 | seizure laterality | | | | | | | | | | | | 1 | | | 9 | | | | | 2 | 9 | 7 | | | 1 | 5 | | 4 | 3 | 2 | 1 | 2 | 9 | | | 4 | 8 | 2 | 6 |
| 3.6 | pre+post epilepsy surgery | | | | | | | | | | | | | | | 1 | | | | | | 2 | 1 | | | 1 | 2 | | 1 | 1 | 1 | | | 2 | | | | 2 | | 1 |
| 3.7 | pre-post drugs epilepsy | | | | | | | | | | | | | | | 1 | | | | | | 1 | | | 1 | 1 | | | 1 | | | | 1 | 1 | 1 | | 1 | 2 | 1 | 2 |
| 3.8 | non-epilepsy drugs | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3.9 | HC/MTL abnormality | | | | | | | | | | | | | | | | | | | | 5 | 16 | 12 | | | 2 | 9 | | 9 | 7 | 4 | 2 | 3 | 15 | 4 | | 6 | 12 | 7 | 9 |
| 3.10 | MCI/SCI | | | | | | | | | | | | | | | | | 1 | | | | 3 | 1 | | | | 2 | | | 1 | | | 1 | 2 | | | | 1 | 2 | 1 |
| 3.11 | Alzheimers/dementia | | | | | | | | | | | | | | | | | | | | | 4 | 5 | | | | | | 4 | 3 | 2 | | 1 | 6 | 4 | | 3 | 3 | 1 | 6 |
| 3.12 | repetitive head impact | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3.13 | nationality/culture | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3.14 | depression/anxiety | | | | | | | | | | | | | | | | | | | | | 7 | 2 | | | 1 | 3 | | 5 | 5 | 1 | | 2 | 6 | 2 | | 2 | 4 | 5 | 3 |
| 4.1 | recall | | | | | | | | | | | | | | | | | | | | | | 17 | | | 2 | 15 | 2 | 20 | 14 | 5 | 1 | 5 | 30 | 7 | 2 | 13 | 21 | 15 | 18 |
| 4.2 | recognition | | | | | | | | | | | | | | | | | | | | | | | | | 1 | 7 | 1 | 10 | 6 | 5 | 2 | 4 | 14 | 5 | | 6 | 13 | 6 | 10 |
| 4.3 | implicit tests | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4.4 | RKG / confidence | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | 1 | 1 | 1 |
| 4.5 | repeated recall/recog | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | 2 | | 1 | | | 3 | 1 | | 2 | 3 | 1 | 3 |
| 5.1 | stories | | | | | | | | | | | | | | | | | | | | | | | | | | | | 6 | 7 | 3 | 2 | 1 | 12 | 2 | | 3 | 9 | 8 | 5 |
| 5.2 | word pairs | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | 2 | | 1 | 1 | 1 | | |
| 5.3 | word Lists | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 11 | 3 | 1 | 2 | 21 | 5 | 1 | 9 | 14 | 10 | 13 |
| 5.4 | visual Figures | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 14 | 3 | | 2 | 10 | 9 | 8 |
| 5.5 | visual Scenes | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 4 | 1 | 1 | 1 | 4 | 1 | 2 |
| 5.6 | routes | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | 1 | 1 | | 1 | 1 | 1 | 1 |
| 5.7 | ABM events anterograde | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 2 | 2 | | 2 | 3 | 5 | 4 |
| 6.1 | 1min-30min | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 7 | 2 | 12 | 23 | 11 | 18 |
| 6.2 | 31m-4hr | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | 4 | 6 | 3 | 6 |
| 6.3 | 4hr-12hr | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 3 | 2 | | 2 |
| 6.4 | 12hr-24hr / nights sleep | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 10 | 2 | 12 |
| 6.5 | 24hr-1week | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 8 | 19 |
| 6.6 | >1week | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 9 |
| 6.7 | multiple (>2) delays | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

A further observation from this initial analysis was the lack of a standardised test for ALF. Each researcher has developed their own customised test, which makes it difficult to compare results across studies and leads to published work being impacted by a range of methodological weaknesses. This also means there is no accepted way to identify or diagnose ALF in a clinical setting. As a result of this observation, a third candidate research area for this PhD was identified:

- development of a test for ALF that overcomes methodological weaknesses in previous work, and can form the basis of a standard test to be used both in research and in clinical practice

## 1.3 Review of literature relevant to selected research areas

Across the lifespan, episodic memory development follows an inverted U-shaped profile with performance increasing to early adulthood, remaining stable during adulthood, and then declining in old age (see Fandakova et al., 2015 for a review). There are multiple models which address the mechanisms leading to the decline seen in later years. One model (Craik & Bialystok, 2006) breaks episodic memory into Control components (flexible goal directed behaviour, use of strategies for encoding and retrieval) and Representation components (crystallised schemas, general knowledge), with evidence that age related decline is mainly due to a decline in the control component (Taconnat et al., 2022). Another model proposes Associative components (mechanisms that bind features of an event) and Strategic components (metacognition, use of strategies), with evidence that age related change is due to a decline in both components (Shing et al., 2010). However, while there is general agreement that memory performance deteriorates in normal ageing, it is less clear whether this reflects purely a reduction in initial acquisition or also entails accelerated forgetting. Studies investigating ageing and forgetting rates have used a range of materials, including images, stories, sentences, word-pairs and lists of words.

Huppert and Kopelman (1989) tested recognition memory for visual images at 10min, 24hr and 1 week. They used multiple methods to control for the initial level of acquisition and found that their older groups displayed faster forgetting, with the greatest difference in rates occurring between 10min and 24hr. In a similar study, Rybarczyk et al. (1987) compared young and older groups recognition memory for line drawings at delays of

10min, 2hr, and 48hr. They matched learning at an individual level through repeated exposure until a criterion was reached and found no evidence of accelerated forgetting at any delay. Park et al. (1986) found evidence of accelerated for images in older participants by 4 weeks, although their use of the same stimuli at all delays may have influenced results, and they made no attempt to match learning. In a follow-up which used unique stimuli at each delay (Park et al., 1988) they found statistically greater forgetting in their older group between 48hrs and 1 week. Baddeley et al. (2014) orally presented short, three or four sentence long descriptions of a series of crimes ('Crimes Test') and then tested memory for different elements of those events at multiple delays using cued-recall. They found evidence of accelerated forgetting in an older group, compared to a young group, at 6 weeks, but not at shorter delays. In a study using a similar paradigm, but enforcing learning to a criterion, Stamate et al. (2022) found their older group forgot faster over 1 month when memory was not refreshed by intermediate testing. Giambra and Arenberg (1993) compared cued-recall memory for sentences in young and older groups at delays up to 24hr. In their first experiment they did not equate initial acquisition between groups and found the initial degree of learning was higher for younger participants, but there was no difference in rates of forgetting. However, in later experiments, when they controlled for initial acquisition level through either adjustment of exposure times or statistically through use of multiple regression they found accelerated forgetting in older participants which was greatest in the first few minutes after learning, then slower but cumulative after that. In a recent, similar study investigating the effect of ageing on cued-recall memory for sentences, but with no attempt to match initial acquisition, Rivera-Lares et al. (2023) found that normal ageing impacted the initial level of acquisition but not forgetting rates to 24hrs. Trahan et al. (1992) tested cued-recall for word-pairs at 30min, without any attempt to match initial learning, and found older participants scored lower at all delays, but showed no accelerated forgetting. Wheeler (2000) found accelerated forgetting for words in older participants after 1 hour, and suggested testing beyond the 20-30mins used in most standardised clinical tests was important to identify differences. However, after using ANCOVA to adjust for initial acquisition, Hulicka et al. (1965) found no evidence of accelerated forgetting for paired-associates after 1 week.

In summary, evidence from studies comparing the forgetting rates of healthy young and older groups is mixed; some studies have found no difference while others have found that older participants have a faster forgetting rate. Differences in methodologies,

including whether initial acquisition was matched, may account for some of the variability between studies and it may be necessary to test beyond the 20-30mins delay used in most standardised clinical tests in order to detect differences.

Of those studies investigating episodic memory performance across the lifespan, testing participants in multiple age bands, many have only done so for memory tested at delays of up to 30mins. In contrast, studies which have addressed ALF in ageing have investigated performance at longer delays, but have often used only two groups, a younger and older group (e.g. Baddeley et al., 2014). While this design is an efficient way to check for an impact of age on memory this does not provide data on the development of any age related decline, including when it starts. However, there are a small number of studies which have tested memory performance both beyond 30mins and across the adult age range.

Davis et al. (2003) tested memory for word lists using multiple age groups (30-45yrs, 46-60yrs, 61-75yrs, 76-90yrs) and at delays of 20min and 24hrs. They found evidence of significant decline in recall at both delays in their two oldest age groups, and presence of accelerated forgetting between 20min and 24hr in the oldest age group. Huppert and Kopelman (1989) tested recognition memory for images at 10min, 24hr and 1 week, and with participants split into 4 age bands (16-19yrs, 19-29yrs, 38-64yrs, 65-83yrs). They found that both older groups displayed accelerated forgetting in the 10min-24hr period, but the width of this time window means it remains unclear whether the forgetting they detected would meet the criteria for ALF. Giambra and Arenberg (1993), in their third experiment, tested cued-recall of sentences at delays up to seven hours, with a participant age range of 18-74yrs. Rather than dividing their participants into age bands they used multiple regression to analyse the impact of age, finding a significant impact on forgetting. However, this analysis does not provide any information on how performance varies across the full lifespan, or when deterioration starts. In summary, the limited extent of the existing literature evaluating memory performance at delays beyond 30mins across the whole adult lifespan indicates a clear need for further research into the lifespan development of memory over longer delays.

ALF has been identified in both healthy older adults, and in those displaying subjective cognitive impairment (SCI) who complain of memory problems but perform normally on standard objective clinical tests. Baddeley et al. (2014) found evidence of ALF in healthy older participants at 6 weeks, but not at shorter delays, using a paradigm based on cued-recall of constrained prose. Manes et al. (2008) found evidence of ALF in older

participants who display SCI. This SCI group were indistinguishable from matched controls at immediate and 30min delays, but were significantly impaired at 6 weeks. Manes et al. suggest that we need more sensitive tests to capture these individuals who complain of memory problems in case they are at risk of developing MCI and perhaps progressing to AD.

Mild cognitive impairment (MCI) is a clinical syndrome reflecting a decline in cognitive function and deficits on objective neuropsychological testing but relatively intact activities of daily living (Dunne et al., 2021; Albert et al., 2011). In contrast, dementia is a general term for a decline in mental ability severe enough to interfere with daily life (McKhann et al., 2011). Alzheimer's disease is the most common cause of dementia. In general practice a doctor may use an easy to administer cognitive screening tool such as MMSE (mini mental state exam, Folstein et al., 1975) or Mini-Cog (Borson et al., 2003) as one factor in identifying those who should be referred to a specialist dementia diagnostic service such as a memory clinic for further investigations which may ultimately lead to a diagnosis of MCI or dementia (National Institute for Health and Care Excellence, 2018). Such a diagnosis requires consideration of multiple factors, which may include a discussion of symptoms with the patient and a family member, a review of medical history, cognitive testing, blood tests and brain scans.

Current empirical evidence linking ALF to both MCI and AD is mixed. For example, with respect to ALF in those diagnosed with AD, Stamate et al. (2020) found only six out of 14 published studies they reviewed had detected evidence for a link, but also highlighted that methodological weaknesses made it difficult to interpret some studies results. In a subsequent experiment of their own they found no evidence of ALF in AD patients once initial learning had been equated using a learning to criterion procedure. They propose that the impairment caused by AD is primarily a learning deficit rather than accelerated forgetting. However, other evidence suggests ALF may be associated with the early stages of the disease process before AD is diagnosed, either before symptoms are observed, or during the early MCI stage. In one of the few studies to investigate ALF in MCI Walsh et al. (2014) tested free recall of short stories at 30min and 1 week. They used a 90% learning criterion, with participants getting additional learning trials until the criterion was reached. They found evidence of accelerated forgetting during the first 30min after learning and between 30min and 1 week. However, when they controlled for the number of trials needed to reach criterion ALF was only observed over the longer 30min-1week delay. This suggests all MCI participants show ALF over the longer delay,

while some also show impaired learning and accelerated forgetting over the short delay. This may reflect a difference in the severity of their condition, and indicate how the disease progresses.

This idea of accelerated forgetting over longer delays being a symptom of an early stage in the AD disease process is backed up by evidence found by Weston et al. (2018) in a study of presymptomatic autosomal dominant AD. This is a familial form of AD, in which carriers of the relevant genetic mutation first display symptoms at relatively predictable ages based on family history. This provides an opportunity to study presymptomatic cognitive change when the remaining time to onset of symptoms can be estimated. Weston et al. found that ALF was detectable in individuals who were on average 7 years away from predicted onset of symptomatic disease, and no learning impairment was observed. They also found a positive correlation between ALF and subjective memory complaints. They interpret their results as showing that ALF is an early presymptomatic feature of autosomal dominant AD, pre-dating other amnestic deficits, and which might underpin subjective memory complaints. If these results generalise to other forms of AD then testing for ALF may be valuable diagnostically and in presymptomatic trials. Further support for this argument comes from the finding that a key genetic risk factor for the more common sporadic late-onset AD, Apolipoprotein (APOE) ε4, is associated with impaired verbal recall and recognition when this is measured at 7 days, but not at 30 minutes (Zimmermann & Butler, 2018). The proposed link between subjective memory complaints and AD is further supported by a meta-analysis by Mitchell et al. (2014) who found that older people with subjective memory complaints who perform normally on objective tests are twice as likely to develop dementia as individuals without complaints. Taken together, this evidence supports the argument that subjective complaints can be the first stage of a progressive decline that moves from subjective memory complaints through MCI to AD.

The possibility of using sensitive cognitive tests for early identification of those at risk of MCI/AD is important given the significant prevalence and impact of the disease and the benefits of early diagnosis. Globally, approximately 50 million people had dementia in 2020. This number is projected to increase to 82 million by 2030 and 152 million by 2050 (World Health Organization, 2020). The impact on the families who often act as care-givers is significant, and can lead to increased emotional and financial stress. The personal, societal and financial implications of AD are therefore profound, and will continue to grow as populations age. Diagnosis at earlier stages allows the individual,

their families and carers to plan for the future, considering elements such as finances and social care plans. It has been estimated that accurate diagnosis of all AD cases at the MCI stage for all individuals alive in the US in 2018 would save $7.9 trillion (Alzheimer's Association, 2018). In addition, Sperling et al. (2011) highlight the need for more sensitive cognitive tests to help track behavioral changes during clinical trials, and to provide a complement to biomarker tests since it seems that a combination of biomarker and cognitive tests may prove most effective in early identification of those at risk. The need for such early diagnosis is expected to grow as the first effective treatments become available. It is expected that these will be most beneficial when intervention happens during the earliest possible stages of the disease process, before clinical symptoms arise, when the maximum cognitive function can be retained. Indeed, once disease-modifying agents are available the risk/benefit ratio of large scale screening may shift towards recommending larger scale screening (Albert, 2007), which will require effective tests that are simple to administer widely.

The cause of ALF remains unclear. Disrupted sleep was a common complaint in many of the early documented cases. As sleep has been found to improve performance on both non-cognitive (Peigneux et al., 2004; Walker et al., 2002) and cognitive (Ellenbogen et al., 2006; Fenn et al., 2003) tasks, there was strong reason to believe that this could be a candidate cause for inefficient memory. Muhlert et al. (2010) discussed the possibility that subclinical epileptiform activity during sleep may be a cause of ALF, as prior to medication TEA patients often report amnestic seizures upon awakening. Using a word-pair learning task Mary et al., (2013) found a correlation between ALF and sleep issues in otherwise healthy older people, suggesting the type of sleep disruption that may cause ALF is not specific to epilepsy.

More recently, however, some studies have demonstrated that even if abnormal activity during sleep *contributes* to certain aspects of memory dysfunction, this may not be the *central* cause of ALF. Atherton et al., (2014) tested TEA patients' memory for unrelated word-pairs after 12 hours of wakefulness or 12 hours that included a night's sleep. Contrary to the sleep-hypothesis, they found that TEA patients benefitted from sleep just as much as matched controls. In fact, even more surprisingly, the patients performed worse during the wakefulness condition. Similarly, Hoefeijzers et al. (2015) studied recall of word lists in a group of TEA patients across the period of a day and found that there was significant forgetting by three hours in the patients.

The inconsistency between different studies is probably driven, at least partly, by methodological differences and weaknesses. In a critical review of existent ALF studies Elliott et al. (2014) highlighted that no standardised clinical anterograde memory tests go beyond 40 minutes. They identified seven requirements for a robust methodology for ALF studies, and highlighted that no existing study met all requirements. As part of this literature review additional methodological weaknesses were collated and added to Elliott et al.'s list. This combined list was then used to produce a detailed set of 17 methodological recommendations for future ALF research, summarised in Table 1.2, where points 1 to 7 are extended versions of the original seven taken from Elliott et al. These recommendations will be used in the design of experiments and tests throughout this PhD.

*Table 1.2 Methodological recommendations for ALF research*

| | |
|---|---|
| 1 | Patient and control groups should be matched, at least for age and intellectual ability, and preferably sex. IQ testing should avoid use of premorbid tests if the condition under study is likely to impact IQ (e.g., epilepsy). |
| 2 | Ideally, both verbal and non-verbal test material should be used. Test difficulty and sensitivity should be equal across material types. In the verbal domain consider ecological validity issues when choosing between word lists/pairs or short stories. |
| 3 | Ideally, forgetting should be measured using both recall and recognition tests. Consider sensitivity if choosing between free recall and cued-recall. |
| 4 | Ceiling and floor effects should be avoided as far as possible. Pilot materials to identify these effects and adjust accordingly. Consider adjusting the difficulty of tasks across delays to limit within participant ceiling/floor effects. Consider adaptive techniques which automatically adjust difficulty on a per-participant basis to limit between participant ceiling/floor effects. |
| 5 | The potential for rehearsal and repeated recall should be avoided as far as possible. Consider use of filled delays, complex stimuli which are difficult to rehearse, decoy stimuli to attract rehearsal efforts away from target stimuli, and use a different set of stimuli for each delay. |
| 6 | The immediate delay period should be long enough to ensure information is stored in LTM and retrieval is not reliant on STM processes. Consider using a distractor task to prevent maintenance on information in STM. |
| 7 | Effort should be made to equate initial learning (whilst avoiding overlearning) to avoid scaling and item difficulty effects. Consider manipulating the presentation time for each stimulus, presenting each stimulus multiple times, and learning to criterion. If manipulating presentation length use fillers to ensure total timings from learning to test are equalised. If learning to criterion do this on a per-item basis to control total recalls per item during learning. If measuring criterion for a full list, rather than per item, consider setting criterion to 80% or similar, rather than 100%, to avoid ceiling effects at short delays. Alternatively do not |

| | |
|---|---|
| | manipulate learning, and instead analyse data using a 'matched sample' approach, matching participants on observed learning. |
| 8 | Effort should be made to measure learning/acquisition rate, and analyse the impact of this on results. Consider comparing sub-groups with similar acquisition rate to eliminate this as a confound. |
| 9 | Control for item difficulty effects by recording individual item recall/recognition rates during pilot studies and either adjust items to equate difficulty, or allow for item difficulty when analysing study results. |
| 10 | If testing with free recall then record per-item measurements so that Subjective Organisation effects can be analysed. |
| 11 | Consider procedures which allow development of interference immunity to be tested. |
| 12 | Record important sleep parameters, including daytime napping, average hours sleep per night, and prevalence of disrupted sleep patterns. |
| 13 | Analyse forgetting rates, not absolute scores. Consider analysis of absolute change in score, percentage change, and analyse from perspective of correct answer count and number of errors. |
| 14 | Measure forgetting at multiple extended delays so that forgetting curves can be quantified. |
| 15 | Carefully analyse forgetting curve shape before statistical significance is reached: does forgetting start early and develop progressively, or start late? |
| 16 | As a general principle recording as much detail as possible during both learning and test, so that previously unidentified factors can be analysed retrospectively. |
| 17 | Procedures and tools should be easily translated to other languages. Where possible develop forms in multiple languages to facilitate cross-cultural and cross-geography generalisation. |

It is impractical for any single test to comply with all 17 recommendations. It is therefore important to focus on those methodological flaws which occur most often and are most likely to impact results.

Ceiling and floor effects (see recommendation 4 in Table 1.2) are common in existing work, and can be expected to have distorted some results. For example, when reviewing existing published work investigating faster forgetting in epilepsy, Cassel et al. (2016) found 11 out of 36 studies showed presence of ceiling effects. In the case of ALF research it is necessary to test performance at a short delay, for example 30mins, and again at a minimum of one long delay. It can be difficult to avoid ceiling effects for high performing groups at short delays while also avoiding floor effects for low performing groups at long delays. Any attempt to develop a standard test for ALF needs to include conscious efforts to overcome this issue.

Elliot et al. (2014) also highlight conditions for learning as an important issue. This is particularly relevant when groups are expected to perform at different levels. For example, in a study of forgetting, if exposure to the stimuli is equal then older groups will generally score lower than younger groups at all time points. In this case, is it valid to directly compare the slopes of the resulting forgetting curves at any given delay? There is ongoing debate in the literature, with some arguing that such comparisons are invalid (e.g. Loftus, 1985) and that they are impacted by confounds such as scaling effects. A loss of one item for a participant scoring at a high level is likely to reflect the loss of a difficult item. In contrast the loss of one item for someone scoring at a low level is likely to reflect the loss of one of the easier items. Does this reflect the same amount of forgetting? The loss of one item for the high performer also reflects less forgetting in relative terms; a loss of one item for someone scoring 10 at the previous timepoint equates to a 10% loss, while a loss of one item for someone scoring 5 equates to 20% loss. Should we compare the relative loss, or the absolute loss? To sidestep these debates it is preferable to equate initial acquisition so that all groups score at the same level at the first test delay.

Some early work on equating initial level of acquisition between groups was performed by Huppert and Piercy (e.g. Huppert & Piercy, 1977; Huppert & Piercy, 1978). They titrated the total duration of exposure to visual images to equate acquisition, such that groups scored similarly at the first test delay. Using this technique they were able to match initial acquisition for organic amnesia patients and healthy controls, and the same approach was used by others to investigate forgetting in a wide range of different patient groups (organic amnesia: Kopelman & Stanhope, 1997; Green & Kopelman, 2002; Isaac & Mayes, 1999a & 1999b; neuropsychiatric disorders: Lewis & Kopelman, 1998; ECT: Squire, 1981; Alzheimer's disease: Kopelman, 1985; Christensen et al., 1998) and in healthy ageing (Huppert & Kopelman, 1989). However, it is important to know whether varying exposure duration in this way introduces any confounds of its own which may influence forgetting.

There is evidence that forgetting curves are independent of the initial degree of learning when this is varied by adjusting exposure duration using the Huppert and Piercy matching procedure. Early work by Slamecka and McElree (1983) and more recent replication and extension by Rivera-Lares et al. (2022) has found that raising the degree of learning of young healthy participants through additional exposure to stimuli did not change the shape of the forgetting curve; instead, different exposures resulted in parallel curves at

different levels. More recently, River-Lares et al. (2023) have shown that the forgetting curves for healthy young and older participants who received the same exposure to stimuli were parallel, despite being at different levels, suggesting that while natural ageing processes may impact the amount of material that is learnt for any given exposure duration, it does not impact forgetting rates. However, all of these studies deliberately avoided use of any testing during the learning procedure. While this matches the methodology used in the Huppert and Piercy matching process and avoids any confounds that could be introduced by testing, it does mean the results cannot be generalised to the type of learning methods used in many other studies, especially much of the recent ALF literature (e.g. Butler at al., 2007), where learning is tested after each presentation of stimuli as part of a 'learning to criterion' procedure.

Learning to criterion is a commonly used method to equate initial acquisition across participants, both to facilitate comparisons of forgetting curves across groups and also to help avoid floor effects at extended delays by ensuring all groups start at a reasonably high level of acquisition. This is typically achieved through increasing the number of learning trials until a particular criterion is reached; for example, 80% correct for free recall of a list of words. This form of learning allows each individual participant's acquisition to be equated, which in addition to equating group means will also reduce within group variance and thus increase statistical power. Testing after each trial also allows the learning rate to be measured, something which is not possible using the Huppert and Piercy paradigm. Irrespective of what level the criterion is set at, if this is applied to the *entire* stimulus set, then the standard procedure is to present the entire set of stimuli repeatedly until the criterion is reached. However, using this procedure, 'quickly learnt' individual items will be successfully recalled more times than others during the learning procedure. It has been shown by Roediger and colleagues (Karpicke & Roediger, 2008; Roediger & Smith, 2012; Roediger & Karpicke, 2006) that multiple successful recalls can confer 'retrieval practice' (see Table 1.2, recommendation 5) and strengthen memories, an effect that would differentially benefit the items that are learnt more quickly, accentuating any initial differences in the learning of individual items, and potentially distorting delayed recall scores. In addition, testing after each learning trial can cause participants to make errors which may also influence later forgetting. Ricci et al. (2015) directly compared forgetting rates for short stories presented once with that for stories presented multiple times and learnt to a criterion of 80%. They found a greater detection of ALF in their epilepsy patients for material which had been learnt to criterion.

Whilst this suggests a learning to criterion process may be more sensitive for detecting ALF, it remains possible that it reflects some level of differential forgetting between groups introduced as an artefact of the learning process through differential retrieval practice and errors.

An additional source of retrieval practice is testing the same material at multiple test delays e.g. words being successfully recalled at one particular time point and then tested again at a later delay. Karpicke & Roediger (2008) have shown that such repeated testing of material significantly improves performance. This means that studies which use the same material at each time point may be masking potential differences between groups at different time points. Baddeley and colleagues tried to avoid this repeated testing of the same material in a series of studies aimed at developing a test for ALF (Baddeley et al., 2014; Baddeley et al., 2019; Baddeley et al., 2021). In their Crimes test participants are read a series of sentences, with each one describing a unique event. Each event is designed such that it has multiple elements which can be probed individually, providing multiple questions for each event, with one question then used at each test delay. This approach was intended to minimize the total learning required while providing enough material to test unique items at each delay. However, Baddeley et al. (2021) found that even partial recall of an event can lead to priming of the non-tested elements of that event, resulting in reduced forgetting for all elements. Using a similar paradigm (Fables test) Stamate et al. (2020) found that patients with Alzheimer's disease benefit from such repeated retrieval just as much as controls. Using a novel story recall paradigm, Jansari et al. (2010) showed that in a patient with TLE, if the patient was repeatedly tested on the same material, his performance was indistinguishable from that of matched controls up to 4 weeks after initial learning. However, if he was tested with a different set of stories at each unique time point, he was significantly impaired within one day and was functionally amnesic to all material within 2 weeks. This study highlighted the impact of repeated testing of the same material potentially masking any underlying forgetting.

The issue of when the accelerated forgetting starts has been debated. In a study aimed at addressing methodological issues highlighted both by Elliott et al. (2014) and themselves, Cassel et al. (2016) tested TLE patients using unique material for each time point to avoid retrieval practice effects. They found that TLE patients forgot both verbal and visual information more rapidly than matched healthy controls, and that the forgetting started early. Contrary to suggestions that ALF is driven by a deficit in secondary consolidation which occurs at extended delays (e.g. Hoefeijzers et al., 2013)

they interpreted their findings as evidence for forgetting starting during the early stage of memory consolidation. In later reviews of TLE studies Cassel and Kopelman (2019) and Mayes et al. (2019) came to differing conclusions. Cassel and Kopelman identified a pattern of early-onset, progressively greater forgetting, and suggest that differences in forgetting patterns reflect a continuum of severity and/or test sensitivity. However, Mayes et al. interpret the existing evidence as supporting the existence of ALF as a qualitatively separate memory condition. To help clarify this, further studies which avoid relevant methodological weaknesses are required.

In work with their TLE patient RY, McGibbon and Jansari (2013) attempted to address a number of the methodological issues highlighted above. In their paradigm, the Verbal Associative Learning and Memory Test (VALMT), they taught unrelated word pairs (e.g. TROOP-SHAWL) to their patient RY and matched controls to a criterion of 100% correct and then used cued-recall (e.g. TROOP-?) to evaluate memory. This paradigm addressed previous methodological weaknesses in the following ways: 1) to address the problem of retrieval practice, learning to criterion was applied at the individual word pair level, rather than being for the whole learning list, with presentation of each word pair ceasing once it had been recalled three times successfully; 2) to address the issue of repeated testing, matched but different word pairs were tested at each of the different time points; 3) to evaluate memory beyond the 40 minutes highlighted by Elliott et al (2014), testing was carried out at four discrete intervals, 5 minutes, 30 minutes, 55 minutes and 4 hours after completion of learning. McGibbon and Jansari found that while their patient was within normal limits at the first two intervals, he was significantly impaired by 55 minutes.

Previous results for short story recall (Jansari et al., 2010) showed RY performed normally at 30 minutes but was impaired at 24 hours. The fact that impairment on story recall tasks was not found until 24 hours can be explained in at least two ways. First, memory had not been tested at an appropriate delay; story recall had only been tested at 30 minutes followed by 24 hours, whereas word-pair cued recall was tested at 30 minutes and 55 minutes. Second, the paradigm used could have driven the difference because the structure of a story may provide enough ‘scaffolding’ to protect a vulnerable memory trace thereby effectively masking forgetting that is already underway. Therefore, although it *is* important to develop more ecologically-valid tests of long-term memory (Baddeley et al., 2014) perhaps for trying to detect accelerated forgetting in the timeframe available to most clinicians (typically one to two hours at most), it will be necessary to develop more

challenging tasks that address the various methodological issues outlined above. If this is the case the VALMT word-pair cued-recall paradigm developed by McGibbon and Jansari (2013) may provide a more sensitive test than that used in many previous ALF studies, and be useful in identifying whether ALF truly reflects late onset forgetting.

There is also reason to think that the paired associative learning used in the VALMT may make it well suited to identifying memory deficits caused by the early stages of AD. Associative learning is vulnerable to the impact of early stage AD (Sapkota et al., 2017) and it relies heavily on hippocampal and entorhinal cortex regions which are known to be vulnerable to change in early AD (de Rover et al., 2011; Coupe et al., 2019; Braak & Braak, 1998).

Overall, there are multiple reasons to believe that the VALMT can form the basis of an ALF test that can be further enhanced and optimised during this PhD and can help answer outstanding questions in the selected research areas: 1) it uses learning to criterion to match learning at a granular level, which has been shown to be effective in the existing literature; 2) it overcomes the differential retrieval practice weakness present in many other learning to criterion techniques; 3) it uses associative learning which is known to be vulnerable to the early stages of Alzheimer's disease, which may make it helpful in identifying those at risk of developing MCI or AD; 4) it has been shown to detect ALF in epilepsy within 55min which means it may be useful for identifying ALF within a single clinical visit; 5) the number of errors made during learning provides a more granular measure of learning performance than the number of trials used in many other learning to criterion techniques; 6) the test format is suited to fully automated online operation which will make is suitable for widescale testing in both research and clinical settings.

## 1.4 Selected research questions

Based on the outcome of the study space analysis and subsequent review of relevant literature, the following specific research problems were selected to form the basis of a PhD research program:

1. Development of the existing VALMT to create an ALF measure which avoids several of the methodological flaws in existing work, and can be used in both research and clinical applications

2. Investigation of ALF prevalence and symptoms in healthy ageing (above 60yrs) and the possibility of using ALF as a marker for those at risk of developing MCI/AD.
3. Investigation of the distribution of ALF, and memory performance at extended delays generally, in the general population, and their relationship to demographic variables such as age
4. Clarification of the timeframe of onset of accelerated forgetting, especially whether forgetting starts early (< 30mins) or late

## 1.5 ALF terminology

The review of existing literature highlighted some differences in terminology between researchers. To avoid any possible misinterpretation, some clarification of the terminology used throughout this thesis is appropriate. First, the initial definition of ALF included normal learning as a requirement; ALF diagnosis using this definition requires learning at a normal rate followed by accelerated forgetting. This eliminates learning performance as a cause of later differences in forgetting rates, and therefore provides the best theoretical justification for the existence of ALF. However, as Baddeley and colleagues (Baddeley et al., 2014; Baddeley et al., 2021; Laverick et al., 2021) suggest, there is no reason to believe that accelerated forgetting at later delays cannot be present in those whose capacity for learning is impaired, and insistence on normal learning performance may have led to under-reporting of ALF. Laverick et al. suggest that use of another term may be appropriate where learning is not normal, for example 'Speeded Long-term Forgetting' (SLF). However, for this thesis the standard term, ALF, will be used, but in the more general sense that includes cases where learning performance is impaired. Second, the term 'accelerated forgetting' will be used to mean a rate of forgetting that is increasing in comparison to that of a control or comparison group. For example, if the forgetting rate for the control group slows over time while the forgetting rate for a patient group remains constant, this would be classed as accelerated forgetting as the difference in group forgetting rates increases with time. A reading of the literature suggests most ALF authors use the term 'accelerated forgetting' in this way.

## 1.6 Statistical analysis

Except where stated otherwise, the following common statistical approach was used for all experiments. Statistical analyses were carried out using SPSS v24 and JASP v0.16.1. Data was checked for normality, homogeneity of variance, and for sphericity as appropriate. Where data was not normally distributed Mann Whitney U (reported as MWU) and Wilcoxon (reported as T) tests were used to compare independent and dependent means respectively. Where the assumption of homogeneity of variance was violated, Welch's F ratio was used and t-tests analyzed assuming unequal variance. Where sphericity was violated, the Greenhouse-Geisser correction for nonsphericity was applied. Overall analyses and interaction effects were investigated using mixed analyses of variance (ANOVA). Comparisons of means across groups at specific delays were performed using one-way ANOVAs and independent samples t-tests. Comparisons of forgetting rates within groups were performed using paired-samples t-tests. All tests were 2-tailed. Bonferroni adjusted LSD post hoc tests were used to investigate significant ANOVA results. Effect size was estimated using $\eta_p^2$, Pearson's $r$ and Cohen's $d$ as appropriate. Where multiple comparisons were conducted using t-tests or correlations, the p-values for any significant results were Bonferroni adjusted by multiplying by the number of comparisons and retaining the standard significance level (p = .05). JASP was used to perform equivalent Bayesian tests where available, using default priors throughout, and the resulting Bayes factors (BF) are reported. Bayes factors quantify the degree to which data support either the null hypothesis (for a 2-tailed test, no effect or group difference exists) or the alternative hypothesis (for a 2-tailed test, an effect or group difference does exist). Bayes factors are reported for the alternative hypothesis throughout, denoted $BF_{10}$. Following the descriptive classifications of Lee and Wagenmakers (2013), $BF_{10}$ Bayes factors between 1 and 3 provide anecdotal evidence for the alternative hypothesis, values between 3 and 10 provide moderate evidence, values between 10 and 30 provide strong evidence, values between 30 and 100 provide very strong evidence, and values above 100 provide extreme evidence. $BF_{10}$ values below 1 provide evidence for the null hypothesis in equivalent bands (anecdotal: 1 to 1/3, moderate: 1/3 to 1/10, strong: 1/10 to 1/30, very strong: 1/30 to 1/100, extreme: smaller than 1/100).

## 1.7 Ethical approval

All research detailed in this thesis was approved by the Research Ethics Board of Goldsmiths, University of London. For all experiments the procedure was explained to participants and their written consent obtained before conducting testing. Participants were informed that they could withdraw from any study at any point without giving a reason and after participation they were provided with a short written debriefing of the research aims and contact details.

# 2 Chapter 2 – Learning and forgetting in healthy ageing

**Note:** Some of the contents of this chapter have been accepted for publication prior to the submission of this thesis, as McGibbon et al. (2022).

## 2.1 Introduction

The literature review summarised in Chapter 1 highlighted the possibility that ALF may occur in healthy ageing, but may also be a symptom of the earliest stages of Alzheimer's disease, while the study space analysis revealed little existing relevant research in either of these areas. There is a clear opportunity to further the state of knowledge in this field, while also addressing a question which may have significant clinical impact. If it is possible to use the forgetting patterns shown by older groups over extended delays to discriminate between healthy ageing and more clinical forms of forgetting, then it may be possible to identify those at risk of developing MCI and AD. This would have considerable benefit, allowing treatments to be provided as early as possible, before disease processes cause significant damage. Even in the absence of effective treatments early diagnosis is valuable, allowing patients and their families to make plans for the future.

This chapter details two studies aimed at investigating whether the accelerated forgetting in healthy older adults documented by researchers such as Mary et al. (2013) and Baddeley et al. (2014), which is usually only assessed at long delays, can be detected at shorter delays using the McGibbon and Jansari (2013) VALMT methodology, and how this would be correlated with subjective memory complaints and sleep quality.

An initial exploratory pilot and follow-up experiment were conducted. The VALMT was used to compare performance of a group of healthy older individuals against that of a group of younger participants, comparing retention at 5, 30 and 55 minutes after first acquiring new information. Results were compared with a standard clinical memory test (WMS-III Logical Memory, Wechsler, 1997), and the relationships to self-reported subjective memory complaints and sleep patterns were investigated.

## 2.2 Experiment 1: Romanian pilot study

As an initial exploratory investigation of forgetting in healthy ageing a pilot study was performed, comparing VALMT performance for a group of 43 Younger participants aged

20-30yrs and a group of 26 Older participants aged 65-80yrs. Participants were recruited and tested in Romania, and for this study the VALMT was translated into Romanian. Memory was tested at delays of 5m, 30m and 55minutes, using a unique set of 12 word-pairs for each delay. As it was not possible to collect any standardised clinical measures of memory or IQ the detailed results of this pilot are not presented. However, the pilot indicated that the test could reveal statistically significant differences in forgetting rates between education-matched younger and healthy older individuals. The overall analysis showed that while the Younger group showed a very shallow forgetting over the period of 55 minutes, the Older group showed a much steeper forgetting function. Importantly, when the Older group were separated based on initial learning rate (number of trials required to reach criterion), different patterns of forgetting were revealed. Older individuals who learned rapidly performed similarly to Younger participants, but those who learned more slowly (to exactly the same criterion) demonstrated lower recall at all delays (5m, 30m and 55m) and a faster rate of forgetting. This relationship with learning performance was therefore examined further in the subsequent main experiment. Full statistics from this pilot are available in Appendix A.

## 2.3 Experiment 2: Forgetting in healthy ageing

### 2.3.1 Rationale

This study was designed to build on the results of Experiment 1 by extending it in four ways. First, the original English version of VALMT, which was designed to investigate ALF in a patient with subclinical temporal lobe epilepsy (McGibbon & Jansari, 2013) was used for the first time to investigate age-related memory decline. Second, building on the previous work indicating a relationship between ALF, subjective memory complaints and progression to AD (Manes et al., 2008; Weston et al., 2018) self-report measures of memory functioning were taken. Third, to address concurrent validity, results from both the VALMT and self-report measures were compared to a standard clinical measure commonly used to assess memory functioning (Wechsler Memory Scale Logical Memory test; Wechsler, 1997). Fourth, following evidence from studies such as Mary et al. (2013) that disrupted sleep can impact memory, sleep quality was explored using a self-report measure.

It was predicted that younger participants would perform better than older participants at delayed recall, that the slow learners amongst the older participants would show

increased evidence of ALF and a higher level of memory complaints, that the VALMT would be a more sensitive measure of ALF than the WMS-LM, and that ALF would be positively correlated with disrupted sleep patterns.

## 2.3.2 Methods

### *2.3.2.1 Participants*

Two groups of participants were assessed: 30 Younger participants aged 19-31yrs (21F, 9M: Mean Age: 24.83, SD: 2.87) were compared to 30 Older participants aged 60-69yrs (20F, 10M: Mean Age: 63.97, SD: 2.54). All participants reported that they were healthy and free from any psychological or medical condition that could have an impact on their memory.

#### 2.3.2.1.1 Demographics and IQ

Groups were matched on gender, education and WTAR predicted IQ (see Table 2.1).

*Table 2.1. Demographic and IQ information as a function of age group*

| **Factor** | **Category** | **Younger *N*(%) or *Mean(SD)*** | **Older *N*(%) or *Mean(SD)*** | **Statistical Test** |
|---|---|---|---|---|
| Gender | Male | 9 (30%) | 10 (33%) | $X^2(1) = 0.08$, $p = .78$, $BF_{10} = 0.35$ |
| | Female | 21(70%) | 20(67%) | |
| Education | Less than High School | 3 (10%) | 4 (13%) | $X^2(3) = 0.74$, $p = .86$, $BF_{10} = 0.09$ |
| | High School | 9 (30%) | 10 (33%) | |
| | Bachelor Degree | 11 (37%) | 8 (27%) | |
| | Graduate Degree | 7 (23%) | 8 (27%) | |
| WTAR[1] Predicted IQ | | 100.23 (5.16) | 100.17 (5.42) | $t(58) = 0.05$, p = .961, $BF_{10} = 0.26$ |

[1] WTAR = Wechsler Test of Adult Reading (WTAR; Holdnack, 2001)

### *2.3.2.2 Materials and measurements*

#### 2.3.2.2.1 Pre-Morbid IQ assessment

All participants completed the Wechsler Test of Adult Reading (WTAR; Holdnack, 2001) to provide a pre-morbid measure of IQ.

#### 2.3.2.2.2 Standard clinical measure of anterograde memory

To provide a comparison for the VALMT and a standard clinical measure of memory, the Wechsler Memory Scale-III UK Edition Logical Memory test, was administered (Wechsler, 1997). The Adult Battery form (Ages 16-69) was used for all groups. To test for immediate and delayed (30mins) recall the Logical Memory I (LMI) and II (LMII) subtests were administered to all participants, respectively. LMI and LMII are composed of the same two stories, each consisting of 25 items. One point was given for each correctly recalled item. The raw scores of the two stories were combined to form a total raw score (max = 50).

#### 2.3.2.2.3 VALMT word-pair cued-recall task

McGibbon and Jansari's (2013) VALMT word pair cued-recall paradigm was used. A subset of the words in the original McGibbon and Jansari (2013) study were used, which were matched for familiarity, concreteness, imageability and frequency (refer to Appendix B for detail). All words were nouns, two syllables and 4-6 letters long. These were used to compose 36 word pairs. Words were randomly assigned to pairs. Words in any pairs with obvious semantic relationships were re-paired.

The stimulus set consisted of three lists of word-pairs (corresponding to the three testing delays under investigation, i.e. 5 mins, 30 mins and 55 mins). Each list consisted of 12 word pairs (e.g. *TROOP* - *SHAWL*). At learning the material from the three lists was interleaved (see Procedure, below). The learning procedure was computer-based, using the same custom written software used in the original 2013 study. Memory was tested using simple paper and pencil.

#### 2.3.2.2.4 Subjective sleep quality

Sleep quality was assessed using the Pittsburgh Sleep Quality Index (Buysse et al., 1989). The PSQI is a self-reported questionnaire composed of nine questions which assess the quality of sleep during the past month across seven domains, with the answers being combined to obtain a global PSQI score for each participant (range 0-21), with a score of five or greater indicating poor sleep quality.

#### 2.3.2.2.5 Subjective memory complaints

To investigate whether participants subjectively report having memory problems, the Memory Complaint Scale (MCS) was administered to all participants (Vale et al., 2012). The questionnaire consists of seven questions. Following standard MCS procedures the scores of all seven questions were summed to form a total MCS score between 0 and 14 and each participant was placed into one of four ordinal categories based on severity of memory complaints: No Memory Complaints (MCs: 0-2), Mild MCs (3-6), Moderate MCs (7-10), Severe MCs (11-14). Finally, to compare data with those of Manes et al (2008), additional custom MCS dichotomous categories were defined: Non-complainers (0-2), Complainers (3-14).

### *2.3.2.3 Procedure*

All testing was performed in a quiet room, either at Goldsmiths, University of London or in the participant's home, with the experimenter as the only other person present.

Participants first completed the WTAR assessment, the self-administered MCS and PSQI questionnaires, and provided demographic information including age and education.

Next the VALMT was administered. To avoid fatigue, the 36 word-pairs were split across three learning periods, with 12 word-pairs learnt during each period. The pairs from the three learning lists were split equally across the three learning periods, so the stimuli for each consisted of 12 pairs total, four from each list (four pairs each from the 5min, 30min and 55min lists; refer to Fig.2.1). Within each learning period the material from the three lists was interleaved so that any changes in strategy, or loss in concentration, or tiredness/stress, would impact all three lists equally.

Initially, each pair (e.g. TROOP - SHAWL) was presented once, for 7s, in a fixed sequence. Once all pairs had been presented, their sequence was randomised. Participants were presented with the first word of a pair (e.g. TROOP-???) and were required to type in the second word of the pair. Immediate feedback was then provided which included display of the correct pairing for two seconds ("Correct. The correct pairing is:…." *or* "Incorrect. The correct pairing is:…."). After displaying all pairs, the process was repeated, using a new random order. This process was repeated in a continual loop with the software having been written such that once any individual pair had been answered

correctly three times it was removed from the list. Once all pairs had been removed from the list (100% learning criterion) the learning session was complete.

Each learning period was followed by a 5 minute rest, and then a test period. During the rest period, participants performed a distraction task (pencil and paper maze completion), to prevent rehearsal. Testing of the relevant material from each learning period was carried out at the three time-points (5mins, 30mins and 55mins); to ensure the correct delays, there were multiple tests (see Fig. 2.1). Where material from multiple lists was being tested at the same time, the pairs were interleaved.

Assessment was carried out using pen and paper; the participant was provided with a sheet that contained the first word of each studied pair, and was asked to fill in the appropriate accompanied word. In total the learning and testing periods, including distractor tasks, lasted 1.5 to 2 hours depending on each individual participant's learning speed.

To avoid interference with the VALMT, all participants were seen again on a different day for administration of the WMS LMI and LMII subtests. During the 30mins rest period between LM1 and LM11 tests participants performed a non-verbal distraction task (pencil and paper maze completion), to prevent rehearsal.

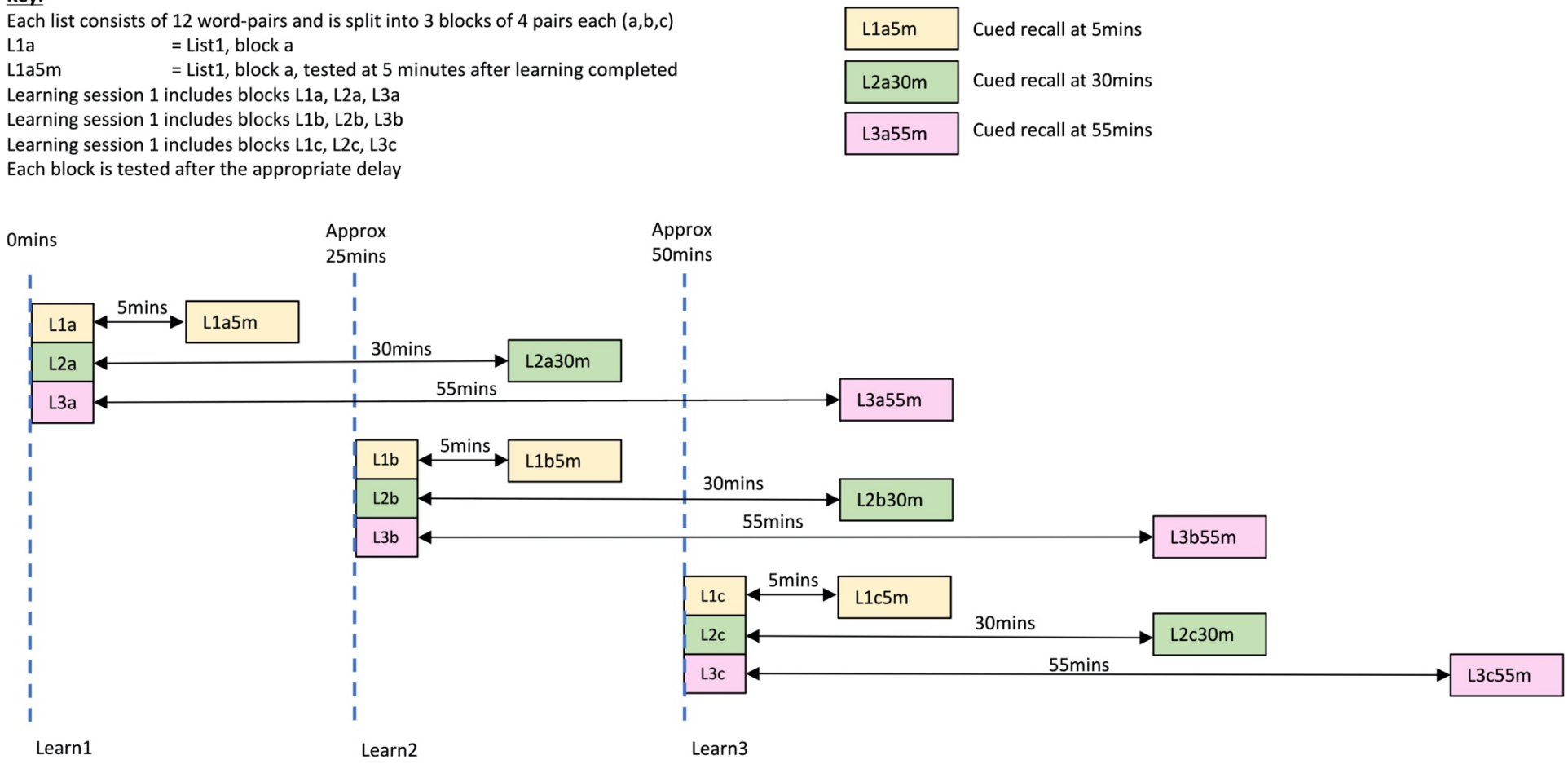


*Figure 2.1 Learning and testing schedule.*

### 2.3.3 Results

#### *2.3.3.1 Learning performance*

A Mann Whitney test was used to compare the number of trials needed to reach criterion (mean across the 3 learning periods). The Older group took significantly more trials to reach criterion ($Mdn_{Older}$ = 56.50 trials, $Mdn_{Younger}$ = 48.50 trials, $MWU$ = 194.5, $p$ < .001, $r$ = 0.57, BF = 52.3).

The variance was significantly greater for the Older group than the Younger group (Older: $SD$=35.4, $\sigma^2$ = 1253.6; Younger: $SD$=10.35, $\sigma^2$ = 107.2; Levene's test for equality of variances violated, $p$ = .005).

To investigate the role of learning performance identified in Experiment 1, the Older group was divided using a median split into those who required fewer trials (Fast_Older, $n$ = 15) and those who required more trials (Slow_Older, $n$ = 15) for some of the later analyses. The split point was set at the median value, 56.5 trials.

The Fast_Older group took more trials than the Younger group to reach criterion, but the difference was small and not significant ($Mdn_{Fast\_Older}$ = 53.00 trials, $Mdn_{Younger}$ = 48.50 trials; $MWU(43)$ = 168.0, $p$ = .17 $r$ = .20, $BF_{10}$ = 0.40), while the Slow_Older group took significantly more trials than the Younger group ($Mdn_{Slow\_Older}$ = 77.00 trials, $Mdn_{Younger}$ = 48.50 trials; $MWU(43)$ = 26.5, Bonferroni adjusted $p$ < .001, $r$ = .71, $BF_{10}$ = 92.63).

A comparison of the two Older groups showed no significant difference in mean age, ($M_{Fast\_Older}$ = 64.07 yrs, $M_{Slow\_Older}$ = 63.87 yrs; $t(28)$ = 0.21, $p$ = .83, $d$ = 0.08, $BF_{10}$ = 0.35), gender ($X^2(1)$ = 0.60, $p$ = .43, $BF_{10}$ = 0.59), education ($X^2(3)$ = 2.9, $p$ = .41, $BF_{10}$ = 0.55), or IQ ($t(28)$ = 0.70, $p$ = .489, d = 0.26, $BF_{10}$ = 0.41).

#### *2.3.3.2 VALMT delayed recall performance*

Figure 2.2 shows the delayed recall performance of the Younger group and the combined Older group, while Figure 2.3 shows the performance for the Fast_Older, Slow_Older and Younger groups.

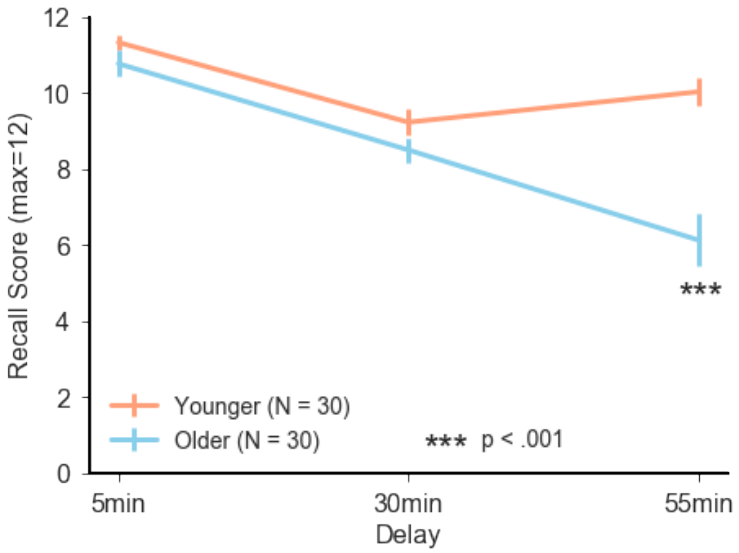


*Figure 2.2 Mean VALMT recall scores as a function of time delay and group (error bars +/- 1SE)*

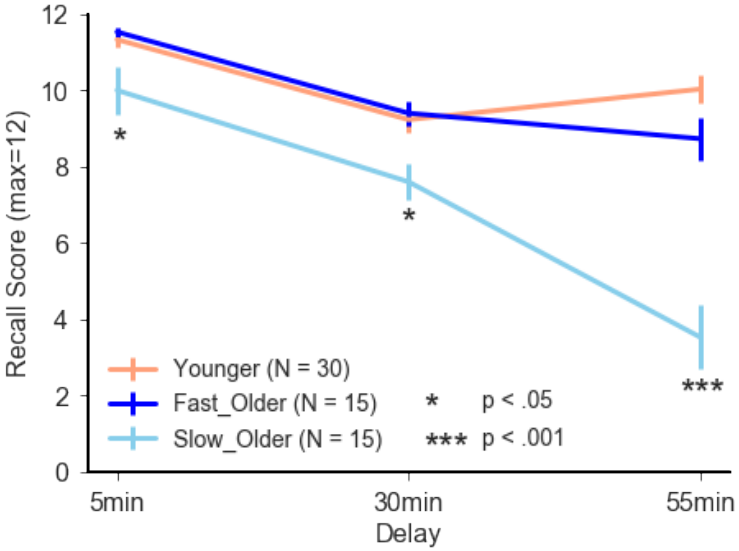


*Figure 2.3 Mean VALMT recall scores as a function time delay and group, separating the Older group into two groups based on initial learning (error bars +/- 1SE)*

#### 2.3.3.2.1 Combined Older group

A mixed factors ANOVA with within-subjects factor Delay (5min vs 30min vs 55min) and between-subjects factor Age (Younger vs Older) was used to analyze cued recall performance across all delay intervals.

There were significant main effects of Age ($F(1, 58) = 13.72$, $p < .001$, $\eta_p^2 = 0.19$, $BF_{10} = 3.00*10^7$) and Delay ($F(1.44, 83.48) = 57.55$, $p < .001$, $\eta_p^2 = 0.50$, $BF_{10} = 2.50*10^{14}$) and a significant interaction between Age and Delay ($F(1.44, 83.48) = 21.47$, $p < .001$, $\eta_p^2 = 0.27$, $BF_{10} = 2.87*10^6$), indicating that compared to the Younger group the

Older group had an overall lower level of recall (Marginal means: $M_{Older}$ = 8.47 pairs, $M_{Younger}$ = 10.20 pairs) and a higher forgetting rate.

To compare cued recall performance across Younger and Older groups at each time-point three independent sample t-tests were used. There was no significant difference at 5mins or 30mins (5mins: $M_{Older}$ = 10.77 pairs, $M_{Younger}$ = 11.33 pairs, $t$ (58) = 1.40, $p$ = .17, $d$ = 0.37, $BF_{10}$ = 0.59; 30mins: $M_{Older}$ = 8.50 pairs, $M_{Younger}$ = 9.23 pairs, $t$ (58) = 1.56, $p$ = .13, $d$ = 0.41, $BF_{10}$ = 0.72). However, the Older group had significantly lower levels of recall compared to the Younger group at 55mins; (55mins: $M_{Older}$ = 6.13 pairs, $M_{Younger}$ = 10.03 pairs, $t$ (43.0) = 5.03, Bonferroni adjusted $p$ <.001, $d$ = 1.32, $BF_{10}$ = 3137).

Forgetting rates were calculated as amount of information lost between two consecutive time points relative to the amount that had been recalled at the earlier of the two time points. Therefore, the 'early' forgetting rate (that between the 5 and 30 minute time points) was calculated as *[5min score – 30min score]/5min score*, and the 'late' forgetting rate (that between the 30 and 55 minute time points) was calculated as *[30min score – 55min score]/30min score*. Independent samples t-tests found no significant difference in early-forgetting rate ($M_{Older}$ = 0.21, $M_{Younger}$ = 0.19; $t$(58) = 0.673, $p$ = .50, $d$ = 0.18, $BF_{10}$ = 0.32), while the Older group had a significantly higher late forgetting rate ($M_{Older}$ = 0.30, $M_{Younger}$ = -0.11; $t$(47.7) = 4.83, Bonferroni adjusted $p$ < .001, $d$ = 1.27, $BF_{10}$ = 1667).

To investigate changes in forgetting rates paired-samples t-tests were used to compare early and late forgetting rates for each group separately. There was a significant difference between early and late rates for the Younger group ($M_{early}$ = 0.187, $M_{late}$ = -0.114; $t$(29) = 4.64, Bonferroni adjusted $p$ < .001, $d$ = 1.53, $BF_{10}$ = 365) but not for Older group ($M_{early}$ = 0.208, $M_{late}$ = 0.301; $t$(29) = 1.21, $p$ = .238, $d$ = 0.32, $BF_{10}$ = 0.37).

#### 2.3.3.2.2 Fast and slow learning Older groups

Equivalent analyses of cued recall performance across all delay intervals were performed with the Older group split into Fast and Slow learning groups using a median split on number of trials required during learning. A mixed factors ANOVA with within-subjects factor Delay (5min vs 30min vs 55min) and between-subjects factor Group (Younger vs Fast_Older vs Slow_Older) identified significant main effects of Delay ($F$(1.6,88.9) = 87.41, $p$ < .001, $\eta_p^2$ = 0.61, $BF_{10}$ = ∞) and Group ($F$(2,57) = 23.22, $p$ < .001, $\eta_p^2$ = 0.45, $BF_{10}$ = ∞), and a significant interaction ($F$(3.1,88.9) = 21.52, $p$ < .001,

$\eta_p^2 = 0.43$, $BF_{10} = 5.51*10^{10}$). Bonferroni adjusted post hoc tests found no significant difference between the Younger and Fast_Older group ($p = 1.00$, $BF_{10} = 0.28$), but significant differences between the Younger and Slow_Older ($p < .001$, $BF_{10} = 1.06*10^7$) and between the Fast_Older and Slow_Older ($p < .001$, $BF_{10} = 1360$).

Recall scores at each individual delay were compared using one-way ANOVAs, and significant results investigated using Bonferroni adjusted post hoc tests for pairwise comparisons. There was a significant difference between the means at all three delays (5mins: $F(2,57) = 5.143$, $p = .009$, $\eta_p^2 = 0.15$, $BF_{10} = 5.08$; 30mins: $F(2,57) = 5.47$, $p = .007$, $\eta_p^2 = 0.16$, $BF_{10} = 6.41$; 55mins: $F(2,57) = 38.42$, $p < .001$, $\eta_p^2 = 0.57$, $BF_{10} = 2.66*10^8$). At all delays the Slow_Older group performed statistically below the Younger (5min: $p = .018$, d = 0.79, $BF_{10} = 3.50$; 30min, $p = .012$, d = 0.88, $BF_{10} = 5.99$; 55min: $p < .001$, d = 2.69, $BF_{10} = 6.67*10^7$) and Fast_Older groups (5min: $p = .019$, d = 0.86, $BF_{10} = 2.61$; 30min, $p = .017$, d = 1.18, $BF_{10} = 12.36$; 55min: $p < .001$, d = 1.89, $BF_{10} = 938$). There was no significant difference between the Younger and Fast_Older groups' performance at any delay (5min: $p = 1.00$, d = 0.20, $BF_{10} = 0.37$; 30min, $p = 1.00$, d = .09, $BF_{10} = 0.32$; 55min: $p = .264$, d = .64, $BF_{10} = 1.54$).

Forgetting rates were compared across groups using one-way ANOVAs, and significant results investigated using Bonferroni adjusted post hoc tests for pairwise comparisons. For early forgetting the difference between the three group means was not significant ($F(2,57) = 0.83$, $p = .44$, $\eta_p^2 = 0.03$, $BF_{10} = 0.26$), whereas the difference was significant for late forgetting ($F(2,57) = 26.20$, $p < .001$, $\eta_p^2 = 0.48$, $BF_{10} = 1.25*10^6$). Post hoc tests comparing the Slow_Older group with the Younger and Fast_Older groups found the Slow_Older had a significantly higher late forgetting rate (Younger: $p < .001$, d = 1.44, $BF_{10} = 6.04*10^5$; Fast_Older: $p < .001$, d = 1.47, $BF_{10} = 67.55$), while there was no significant difference between the Younger and Fast_Older groups' late forgetting rates ($p = .174$, d = 0.21, $BF_{10} = 2.18$).

To investigate changes in forgetting rates for the older groups paired-samples t-tests were used to compare early and late forgetting rates for each group separately. There was no significant difference between early and late rate for the Fast_Older group ($M_{early} = 0.184$, $M_{late} = 0.062$; $t(18) = 1.57$, $p = .14$, d = 0.64, $BF_{10} = 0.72$). However, the difference was significant for the Slow_Older group ($M_{early} = 0.23$, $M_{late} = 0.54$; $t(14) = 2.83$, Bonferroni adjusted $p = .026$, $d = 1.11$, $BF_{10} = 4.41$).

#### 2.3.3.2.3 Relationship between learning period and recall at 55min delay

The VALMT procedure used in Experiment 1 and in the current experiment includes interleaving of learning and testing. This ensures that any changes in strategy, or loss in concentration, or fatigue or stress will impact the learning of all lists equally, and makes it feasible to complete the entire process within a single visit. However, it has the potential to create interference occurring between learning and test that may negatively impact delayed recall. Any such effect would be greatest for the pairs tested at 55min. Referring to Figure 2.1, for these 55min pairs, the items learnt in the first learning period, learning period 1, and then tested after 55min (L3a pairs) have the greatest potential to experience interference, while those learnt in learning period 2 (L3b pairs) should encounter less interference, and finally those learnt in learning period 3 (L3c pairs) should encounter the least interference. Using this rationale, if interference was impacting results then an inverse pattern in recall scores would be expected, with L3a pairs getting the lowest recall score, L3b pairs the second highest and L3c pairs the highest. The mean recall scores for each of these 3 sets are summarized in Table 2.2.

*Table 2.2 55min delayed recall as a function of group and period in which pairs were learnt, separating the Older group into two groups based on initial learning.*

| Group | Learning Period 1 (L3a) Mean (SD) | Learning Period 2 (L3b) Mean (SD) | Learning Period 3 (L3c) Mean (SD) |
|---|---|---|---|
| Younger (N=30) | 3.00 (1.14) | 3.47 (0.90) | 3.57 (0.82) |
| Fast_Older (N=15) | 2.60 (1.12) | 3.13 (1.25) | 3.00 (0.93) |
| Slow_Older (N=15) | 1.20 (1.15) | 1.27 (1.38) | 1.07 (1.16) |

Applying a mixed factors ANOVA with within-subjects factor Learning Period (learning period 1[L3a] vs learning period 2[L3b] vs learning period 3[L3c]) and between-subjects factor Group (Younger vs Fast_Older vs Slow_Older) identified a significant main effect of Group ($F(2,57) = 38.43$, $p < .001$, $\eta_p^2 = 0.57$, $BF_{10} = 2.27*10^8$), but no significant effect of Learning Period ($F(2,114) = 2.39$, $p = .10$, $\eta_p^2 = 0.04$, $BF_{10} = 0.86$) and no significant interaction ($F(4,114) = .86$, $p = .49$, $\eta_p^2 = 0.03$, $BF_{10} = 0.28$). While no significant effect of Learning Period was found, the Bayes factor indicates the data provide only anecdotal evidence for a lack of effect (the null hypothesis). However, the Bayes factor for the interaction shows strong evidence for the lack of an interaction, indicating that interference does not impact groups differently and therefore is unlikely to be the cause of large observed group differences in recall.

#### *2.3.3.3 WMS Logical Memory story recall performance*

##### 2.3.3.3.1 Standardised assessment

All participants scores were compared to the WMS-LM normative data, which showed that none of the participants were impaired.

##### 2.3.3.3.2 Recall performance

Figure 2.4 shows the free recall performance of the Younger group and the combined Older group, while Figure 2.5 shows the performance for the Younger, Fast_Older and Slow_Older groups (please note that these groupings are based on VALMT learning performance as above).

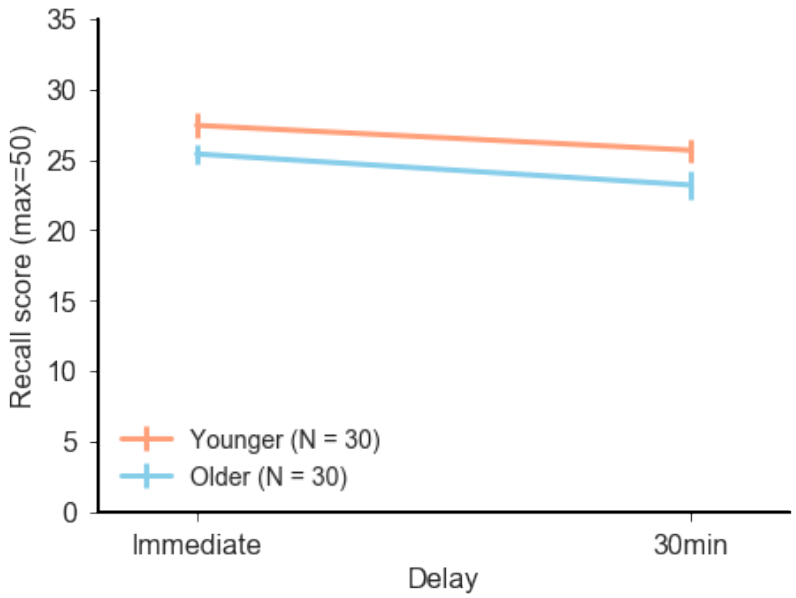


*Figure 2.4 Mean WMS-LM recall scores as a function of time delay and group (error bars +/- 1SE)*

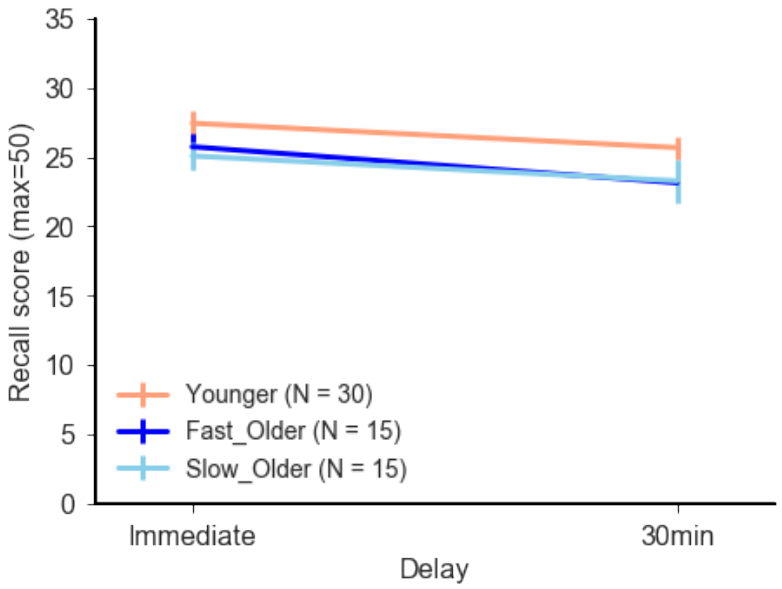


*Figure 2.5 Mean WMS-LM recall scores as a function time delay and group, separating the Older group into two groups based on initial learning (error bars +/- 1SE)*

#### 2.3.3.3.2.1 Combined Older group

A mixed factors ANOVA with within-subjects factor Delay (immediate vs 30min) and between-subjects factor Age (Younger vs Older) was used to analyze cued recall performance across all delay intervals.

There was a significant main effect of Delay ($F(1, 58) = 18.49$, $p < .001$, $\eta_p^2 = 0.24$, $BF_{10} = 272$). The main effect of Age approached significance ($F(1, 58) = 4.01$, $p = .05$ $\eta_p^2 = 0.065$, $BF_{10} = 1.20$), and there was no significant interaction between Age and Delay ($F(1, 58) = 0.22$, $p = .64$, $\eta_p^2 = .004$, $BF_{10} = 0.63$), indicating that compared to the Younger group the Older group had an overall lower level of recall (Marginal means: $M_{Older} = 24.3$, $M_{Young} = 26.5$) but did not differ in forgetting rate.

To compare recall performance across Younger and Older groups at each time-point two independent sample t-tests were used. There was no significant difference in immediate or 30mins recall (immediate: $M_{Older} = 25.4$, $M_{Younger} = 27.4$, $t(58) = 1.81$, $p = .075$, $d = 0.48$, $BF_{10} = 1.03$; 30mins: $M_{Older} = 23.2$, $M_{Younger} = 25.7$, $t(58) = 1.89$, $p = .063$, $d = 0.50$, $BF_{10} = 1.17$).

Forgetting rates were calculated as [immediate recall – 30m recall)]/ immediate recall. An independent samples t-test found no significant difference in forgetting rate ($M_{Older} = .085$, $M_{Younger} = .060$; $t(58) = 0.634$, $p = .53$, $d = 0.17$, $BF_{10} = 0.31$).

#### 2.3.3.3.2.2 Fast and slow learning Older groups

Equivalent analyses were performed with the Older group split into the Fast and Slow learning groups identified earlier. A mixed factors ANOVA with within-subjects factor Delay (immediate vs 30min) and between-subjects factor Group (Younger vs Fast_Older vs Slow_Older) identified a significant main effect of Delay ($F(1,57) = 17.67$, $p < .001$, $\eta_p^2 = 0.24$, $BF_{10} = 257$), but no significant main effect of Group ($F(2,57) = 1.99$, $p = .15$, $\eta_p^2 = 0.065$, $BF_{10} = 0.59$) and no significant interaction ($F(2,57) = 0.29$, $p = .75$, $\eta_p^2 = 0.01$, $BF_{10} = 0.28$).

Recall scores at each delay were compared using one-way ANOVAs. There was no significant difference between the means at either delay (immediate: $F(2,57) = 1.71$, $p = .191$, $\eta_p^2 = 0.06$, $BF_{10} = 0.51$; 30mins: $F(2,57) = 1.77$, $p = .179$, $\eta_p^2 = 0.06$, $BF_{10} = 0.54$).

Forgetting rates were compared across groups using a one-way ANOVA, with no significant difference found ($F(2,57) = 0.40$, $p = .673$, $\eta_p^2 = 0.01$, $BF_{10} = 0.19$).

### *2.3.3.4 Subjective sleep quality*

Figure 2.6 shows the sleep quality scores of the Younger group, the combined Older group, and the Older group split into Fast_Older and Slow_Older as described above.

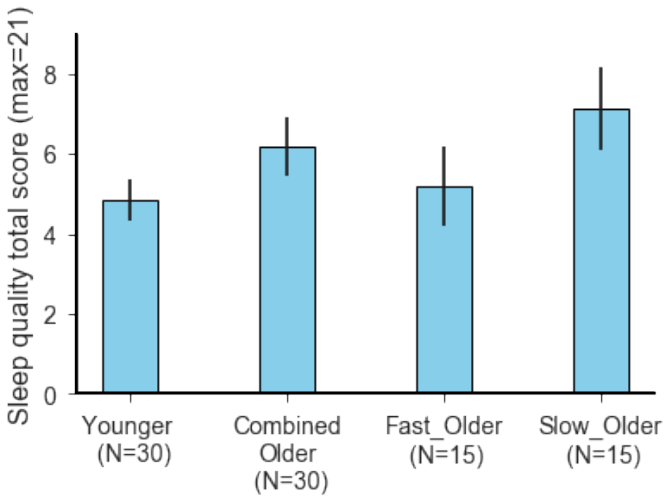


*Figure 2.6 Sleep quality total score as a function of Group, with Older participants shown as a combined group and separated into two groups based on initial learning (error bars +/- 1SE)*

#### 2.3.3.4.1 Combined Older group

Sleep quality scores for Younger and Older groups were compared using an independent sample t-test. There was no significant difference ($M_{Older}$ = 6.17, $M_{Younger}$ = 4.83, $t(52.39) = 1.49$, $p = .14$, $d = 0.39$, $BF_{10} = 0.67$).

#### 2.3.3.4.2 Fast and slow learning Older groups

Sleep quality scores were compared using a one-way ANOVA. There was no significant difference between the means of the Younger, Fast_Older and Slow_Older groups ($M_{Younger}$ = 4.83, $M_{Fast\_Older}$ = 5.20, $M_{Slow\_Older}$ = 7.13, $F(2,57) = 2.35$, $p = .104$, $\eta_p^2 = 0.08$, $BF_{10} = 0.77$).

### *2.3.3.5 Subjective memory complaints*

Figure 2.7 shows the total MCS (Memory Complaint Scale) scores of the Younger group, the combined Older group, and the Older group split into Fast_Older and Slow_Older groups.

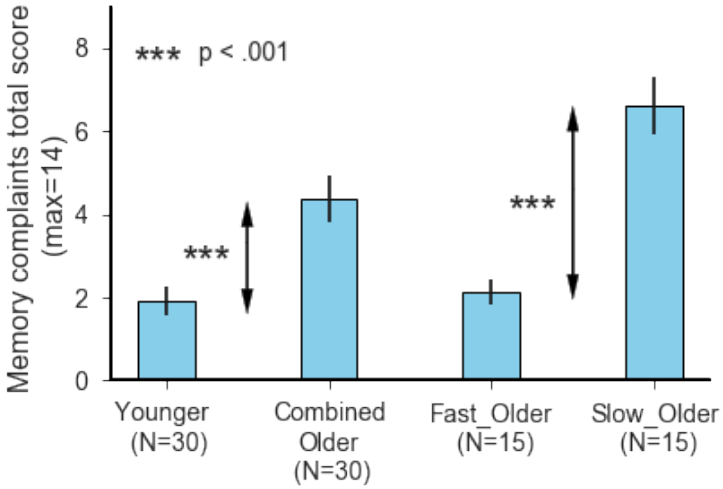


*Figure 2.7 Memory complaints total score as a function of Group, with Older participants shown as a combined group and separated into two groups based on initial learning (error bars +/- 1SE)*

#### 2.3.3.5.1 Combined Older group

Memory complaint scores for Younger and Older groups were compared using an independent sample t-test. The Older group had significantly higher complaints ($M_{Older}$ = 4.37, $M_{Younger}$ = 1.93, $t$ (48.6) = 3.73, $p = .001$, $d = 0.98$, $BF_{10} = 61.79$).

#### 2.3.3.5.2 Fast and slow learning Older groups

Memory complaint scores were compared using a one-way ANOVA. There was a significant effect of Group on memory complaints ($F(2,57) = 30.72$, $p < .001$, $\eta_p^2 = 0.52$, $BF_{10} = 9.21*10^6$). Bonferroni adjusted post hoc tests confirmed there was no significant difference between the Younger and Fast_Older groups (Younger vs Fast_Older $p = 1.00$, $d = 0.11$, $BF_{10} = 0.33$), while there was a significant difference between the Slow_Older and both the Younger and Fast_Older groups (Younger vs Slow_Older $p < .001$, $d = 2.16$, $BF_{10} = 3.74*10^5$; Fast_Older vs Slow_Older $p < .001$, $d = 2.18$, $BF_{10} = 6514$).

Table 2.3 summarises the MCS scores for each group when categorised into four ordinal categories (see Methods). 80% of Younger and 67% of Fast_Older were non-complainers, whereas 87% of the Slow_Older were complainers (20% Mild and 67% Moderate complainers).

*Table 2.3 Distribution of Memory Complaints across Age Groups*

| **Memory Complaints Category** | **Younger *N*(%)** | **Fast_older *N*(%)** | **Slow_older *N*(%)** |
|---|---|---|---|
| None | 24 (80%) | 10 (67%) | 2 (13%) |
| Mild | 4 (13%) | 5 (33%) | 3 (20%) |
| Moderate | 2 (7%) | 0 (0%) | 10 (67%) |
| Severe | 0 (0%) | 0 (0%) | 0 (0%) |

#### 2.3.3.5.3 Relationship between subjective memory complaints and VALMT scores

There was a significant correlation between memory complaints and VALMT recall score at 55mins for the Older group ($r = -0.74$, Bonferroni adjusted $p<0.001$, $BF_{10} = 7427$) but not for the Younger group ($r = -.09$, $p=.62$, $BF_{10} = 0.26$).

Irrespective of initial learning performance (fast or slow), the Older group were now separated into those who reported no subjective memory complaints and those who reported complaints (combining Mild, Moderate and Severe categories), and VALMT performance for these two groups was analysed. Figure 2.8 shows the VALMT recall scores for the Non-complainers and Complainers.

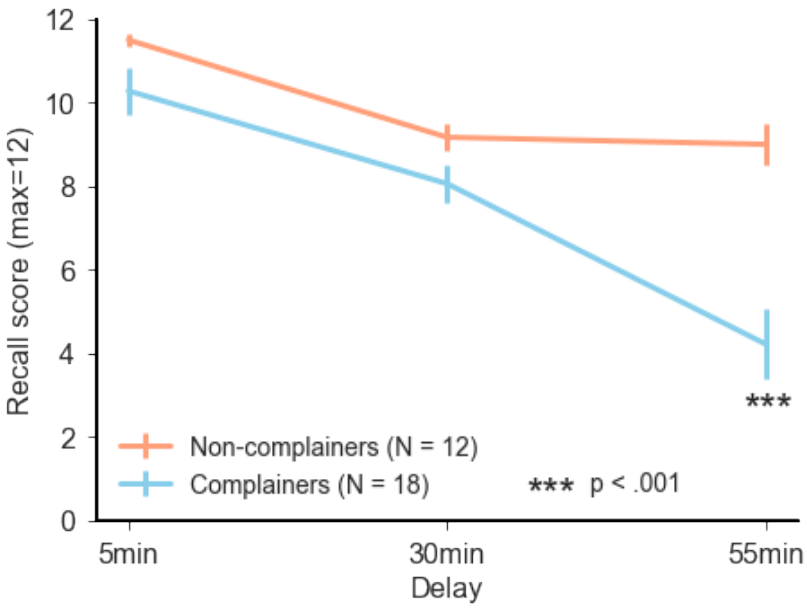


*Figure 2.8 Mean VALMT recall scores as a function time delay and Group, separating the Older group into two groups based on memory complaints (error bars +/- 1SE)*

A mixed factors ANOVA with within-subjects factor Delay (5min vs 30min vs 55min) and between-subjects factor Group (Non-Complainers vs Complainers) was used to analyze cued recall performance across all delay intervals.

There were significant main effects of Delay ($F(1.32, 37.0) = 47.99$, $p < .001$, $\eta_p^2 = 0.63$, $BF_{10} = 2.48*10^{12}$) and Group ($F(1, 28) = 12.48$, $p = .001$, $\eta_p^2 = 0.31$, $BF_{10} = 5048$) and a significant interaction between Delay and Group ($F(1.32, 37.0) = 11.39$, $p < .001$, $\eta_p^2 = .29$, $BF_{10} = 1314$), indicating that compared to the Non-complainers, the Complainers had an overall lower level of recall (Marginal means: $M_{\text{Non-complainers}} = 9.89$ pairs, $M_{\text{Complainers}} = 7.52$ pairs) and a higher forgetting rate.

To compare cued recall performance across Non-complainers and Complainers at each time-point, three independent sample t-tests were used. The difference was significant at 55mins, not significant at 30mins, and at 5mins it was significant before but not after adjusting for multiple comparisons (5mins: $M_{\text{Non-complainers}} = 11.5$ pairs, $M_{\text{Complainers}} = 10.3$ pairs, $t(19.5) = 2.15$, $p = .04$, Bonferroni adjusted p = 0.12, $d = 0.68$, $BF_{10} = 1.12$; 30mins: $M_{\text{Non-complainers}} = 9.17$ pairs, $M_{\text{Complainers}} = 8.06$ pairs, $t(28) = 1.76$, $p = .09$, $d = 0.68$, $BF_{10} = 1.09$; 55mins: $M_{\text{Non-complainers}} = 9.0$ pairs, $M_{\text{Complainers}} = 4.22$ pairs, $t(25.9) = 4.88$, Bonferroni adjusted $p < .001$, $d = 1.64$, $BF_{10} = 117$).

Independent samples t-tests found no significant difference in early-forgetting rate between the two groups ($M_{\text{Non-complainers}} = 0.20$, $M_{\text{Complainers}} = 0.21$; $t(28) = 0.301$, $p = .77$, $d = 0.12$, $BF_{10} = 0.36$), but the Complainers had a significantly higher late-forgetting rate

($M_{\text{Non-complainers}}$ = 0.01, $M_{\text{Complainers}}$ = 0.50; $t(28) = 4.04$, Bonferroni adjusted $p < .001$, $d$ = 1.55, $BF_{10} = 67.79$).

#### 2.3.3.5.4 Relationship between subjective memory complaints and WMS LM scores

Figure 2.9 shows the WMS LM recall scores for the Older group, when split into Non-complainers and Complainers.

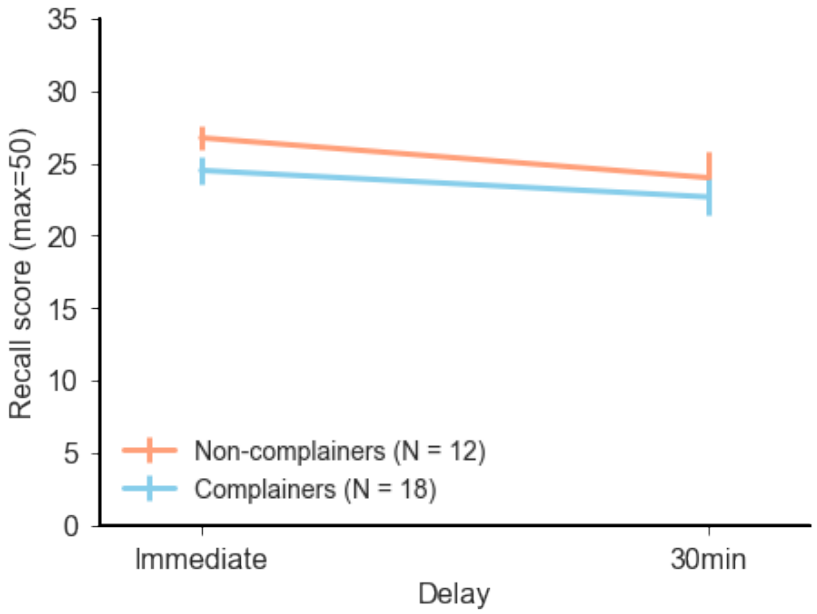


*Figure 2.9 Mean WMS-LM recall scores as a function time delay and Group, separating the Older group into two groups based on memory complaints (error bars +/- 1SE)*

A mixed factors ANOVA with within-subjects factor Delay (immediate vs 30min) and between-subjects factor Group (Non-Complainers vs Complainers) was used to analyze cued recall performance across all delay intervals.

There was a significant main effect of Delay ($F(1, 28) = 7.50$, $p = .01$, $\eta_p^2 = 0.21$, $BF_{10}$ = 3.44) but no significant effect of Group ($F(1, 28) = 1.33$, $p = .26$, $\eta_p^2 = 0.04$, $BF_{10}$ = 0.55) and no significant interaction between Delay and Group ($F(1, 28) = 0.30$, $p = .59$, $\eta_p^2 = .01$, $BF_{10} = 0.50$).

#### 2.3.3.5.5 Relationship between the Older groups learning performance and subjective memory complaints, response-time, and age

To provide an indication of familiarity with using computers the average time taken to answer each cued-recall query during learning was calculated ('*response-time*'). The rationale was that participants unfamiliar with computers may also take longer to type

their answers, and/or spend more time thinking about what to do. The mean number of trials needed to reach criterion was then correlated with *response-time*, total MCS scores and age, adjusting for multiple comparisons by multiplying the p-values by number of comparisons (3).

The correlation of trials to criterion with total MCS scores was large and significant (rho = .554, adjusted p = .006), while the correlation with *response-time* was very small and not significant (rho = 0.041, adjusted p = 1.0) and the correlation with age was small and not significant (rho = -.157, adjusted p = 1.0).

### 2.3.4 Discussion

An initial exploratory experiment (Experiment 1) demonstrated that the VALMT word-pair learning paradigm developed by McGibbon and Jansari (2013) can reveal differences in forgetting rates between education matched, healthy younger and older individuals within a 55 minute delay after learning. As with others (e.g. Davis, 2003; Morse, 1993) a larger variation was found in performance in the Older group than the Younger group, in both learning and recall performance. The greater forgetting seen in the Older group was largely driven by those who required more trials to learn to criterion, or 'slow learners,' (Slow_Older), while those who required fewer trials, or 'fast learners', (Fast_Older) performed very similarly to the Younger group.

Experiment 2 built on these findings, adding additional experimental controls, adding measures for subjective memory complaints and sleep quality, adding a standardised anterograde memory test for comparison, and matching for IQ. In agreement with predictions and the results from Experiment 1, the overall analysis showed that while the Younger group showed a very shallow forgetting over the period of 55 minutes, the combined Older group showed a steeper forgetting function, with the difference in performance reaching statistical significance by 55mins. When the Older group was separated based on initial learning rate (number of trials to reach criterion), different patterns of forgetting were again revealed. As in Experiment 1, the fast learning Fast_Older group performed similarly to the Younger group at all delays, while the slow learning Slow_Older group demonstrated lower recall at all time points, and a faster rate of forgetting between 30 and 55mins. Importantly, performance on a standardised assessment (WMS-LM) suggested that none of these healthy participants had a memory impairment that would be diagnosed using this existing clinical measure.

As the VALMT is a new paradigm, it is important to understand how sensitive it is in comparison to existing standard clinical measures. To evaluate, this a comparison with the WMS Logical Memory test was performed. As the WMS-LM measures recall immediately after learning and after a 30min delay it was not possible to compare performance with the VALMT at the longer 55min delay. Over the initial 30mins the WMS-LM results for the combined Older group matched the VALMT results: no statistical difference was found between Younger and Older groups in either recall performance or forgetting rate. However, once the Older group was split into Fast and Slow learners a clear difference appeared. At 30mins WMS-LM was unable to identify any difference between the Younger, Fast_Older and Slow_Older groups, while the VALMT identified lower recall performance for the Slow_Older group, indicating that the VALMT is able to reveal rapid forgetting in a manner that a standard clinical measure is unable to. The VALMT 55min results show a further large drop in Slow_Older performance and a significantly accelerated late forgetting rate (30-55min).

Group differences in self-reported sleep quality were small, and not statistically significant. However, they did follow the same ordinal pattern as recall performance, with the Slow_Older group showing the worst sleep quality, followed by Fast_Older and then Younger groups. One possible explanation for the lack of a significant difference may be the slightly younger age of the Older group (60-69yrs) in comparison to those in other studies such as Mary et al. (2013; 65-75yrs).

In contrast, the group differences in self-reported memory complaints were large, and statistically significant. The combined Older group reported more memory complaints than the Younger group, with a large effect size. When the Older group was split into Fast and Slow learners, the analysis showed that the bulk of the difference in memory complaints was driven by the Slow_Older group. When the combined Older group was split into Complainers and Non-complainers there was a clear difference in VALMT performance. The Complainers scored significantly lower at 55min, and had a higher late forgetting rate (30-55min). This provides strong evidence of a link between subjective memory complaints and objective cued-recall performance as measured by the VALMT. The WMS-LM, in contrast, was unable to differentiate between these groups, with no difference in recall scores or forgetting rate, although it should be noted the longest delay for this test was 30min, not 55min. This suggests the VALMT may be better able to identify subtle objective memory differences that are linked to subjective complaints,

although a comprehensive comparison against WMS-LM will require future testing at 55min delay.

#### *2.3.4.1 Why were differences in performance identified at shorter timeframes than previous studies?*

How is it that healthy older individuals who successfully learnt material to criterion, and were not impaired on a standard clinical measure of memory, show significant impairment on the VALMT within 55 minutes?

Although no participants reported any diagnosed psychological or medical conditions that affect memory perhaps some have an undiagnosed medical condition of this type, in particular, Alzheimer's disease or another form of dementia. As it was not possible to administer a standardised neuropsychological test battery this possibility cannot be ruled out. However, episodic memory is generally the first mental capacity to be impacted by such conditions, and therefore the fact that all participants performed normally on the WMS-LM, a standard test of anterograde memory, argues against this.

In the Manes et al. (2008) study, the older 'Complaint' group who showed accelerated forgetting at 6 weeks were unimpaired at 30 minutes on their story recall task This mirrors the pattern seen in the WMS-LM results from this experiment, where no difference was found between the groups at immediate recall or 30mins. It may be that story-recall performance would differ between groups if tested at the longer 55min delay. However, a more likely explanation seems to be the type of task.

Due to the 'scaffolding' that story grammars provide, recalling such material is generally an easier task than remembering word pairs that are unrelated. This may well have contributed to the normal performance of the Manes et al. (2008) Complaint group at 30 minutes. Indeed, McGibbon and Jansari (2013) have demonstrated that RY, their patient with subclinical epilepsy, could pass the standard tests of memory as well as a story recall task at 30 minutes (Jansari et al, 2010) but still show a significant impairment on their word-pair learning task by 55mins. Further evidence for this argument comes from the comparison of VALMT and WMS-LM scores, with the story recall based WMS-LM showing no group differences, while the VALMT identified significant differences in both recall level at all time points (5, 30 and 55mins) and late forgetting rate (30-55m).

In common with Manes et al. (2008), Baddeley et al. (2014) did not find a significant difference until 6 weeks when comparing younger and older participants. Their task used

cued-recall of events, each detailed in 3 or 4 sentences ('constrained prose'; 'Crimes Test'). At each delay, they tested recall for a separate subset of items from each event, to avoid repeated recall. Although the test involved answering questions about each event rather than freely recalling it, the argument that 'scaffolding' with integrated material makes the task easier, would apply. Although they did test at intermediate intervals varying between 24hrs and 24 days they failed to find any significant differences at these timepoints. However, only eight participants were tested at each intermediate delay, which they acknowledge may have contributed to the failure to find significant forgetting at these shorter delays. Furthermore, in later work Baddeley et al. (2019) found evidence that although their Crimes Test procedure evaluates recall of different elements of each event at each delay (to avoid repeated recall), this partial recall was priming the other elements of each event, resulting in reduced forgetting of these non-tested elements. In follow-up work using a modified procedure with recall of separate unrelated events at each delay (Baddeley et al., 2021), they found this priming was avoided and forgetting was more rapid. It is possible that using this modified procedure might lead to detection of ALF in older populations at shorter delays than 6 weeks.

In addition to facilitating delayed recall, the integrated nature of materials such as short stories may also hide learning deficits in older participants. Naveh-Benjamin et al. (2003) show that ageing impacts formation of associations between arbitrary items (their 'Associative Deficit Hypothesis', ADH), and that this associative deficit is reduced when the components of the episode are already connected in memory, such as is the case with coherent stories or semantically related words. In contrast, the use of semantically unrelated word-pairs, as in the VALMT, would be expected to highlight any such deficit.

Baddeley et al. (2014) rightly point out that story recall is a more naturalistic paradigm which is useful for looking at long-term memory issues. However, from a clinical perspective it is desirable to obtain objective evidence within a limited amount of time, preferably within a single clinical session, and given the arguments above, this can be difficult with story recall. In contrast, importantly, results from Experiments 1 and 2 suggest it may be possible to obtain such evidence within a single visit using the VALMT.

In the Mary et al. (2013) study, the older group were also unimpaired at the 30m interval on their task, which like VALMT was also a word-pair learning task. However, in their paradigm, to help participants develop associations between the unrelated words, diagrams depicting the words were presented and participants were actively encouraged

to use mental imagery to associate the words together. Further, the older participants were exposed to the stimuli for twice as long the younger participants. Since both mental imagery (Hussey et al., 2012; Sheldon et al., 2017) and increased exposure times to materials (e.g. Ganor-Stern et al., 1998) are known to improve retention, these factors will have strengthened memory for the older group, which may have masked any difference between groups at the 30 minute interval.

A further possible explanation for the early differences identified in delayed recall performance may be interference; specifically, differences in the amount of interference each group is exposed to, or differences in groups' vulnerability to interference, or a combination of these factors. The VALMT procedure can introduce interference in two main ways.

First, interference may be introduced within a single learning period (learning a set of 12 pairs). Once a pair has been learnt to criterion it stops being re-presented, while testing and presentation of the remaining pairs continues. This may create retroactive interference towards the previously learnt pairs. In addition, the errors made during unsuccessful recalls of a given pair may create interference for that pair. Indeed, there is some evidence that using procedures that eliminate errors during learning ('errorless learning') can be beneficial for healthy elderly (Baddeley & Wilson, 1994; Wilson et al., 1994) and effective as a memory rehabilitation technique for AD patients (Clare et al., 2002). Both of these potential sources of interference within a learning period will be worse for slow learners who require more trials overall. Note that a 'trial' in VALMT refers to an attempt to recall a single pair, whereas a trial in most other paradigms refers to an attempt to recall an entire list of words.

Second, the overall VALMT procedure used in Experiments 1 and 2 is relatively complex and involves interleaving of multiple learning and testing phases. The additional learning and testing activities that occur between learning and delayed test of a given pair may create interference. Any such interference will be greatest for material recalled at 55min and least for material recalled at 5min, and again will be worse for slow learners who require more trials over all.

To empirically investigate the effects of interference the delayed recall results for pairs recalled at 55min pairs were broken down by learning period. One third of these pairs are learnt during learning period 1 (L3a), one third during learning period 2 (L3b) and one third during learning period 3 (L3c). Potential interference should be greatest for those learnt during learning period 1, intermediate for those learnt during period 2, and

smallest for those learnt during period 3 (refer to Figure 2.1 for detail). By comparing recall results for these subsets it was possible to investigate the impact of interference induced by the procedure. If interference was impacting results L3a would be expected to get the lowest recall score (as it encounters the greatest potential interference), L3b the second highest and L3c the highest score. In fact, there was no statistically significant main effect of learning period, although the Bayes factor indicates the data provide only anecdotal evidence for a lack of effect. However, there was strong evidence for a lack of an interaction between learning period and group (Younger, Fast Older, Slow Older), indicating that any interference does not impact groups differently. Taken together this suggests inconclusive evidence that interference is not a cause of forgetting between delays, and strong evidence that it is not the cause of the observed group differences in forgetting.

While this data argues against interference as a primary driver of the observed results, further work will be required to fully validate this. For example, future work could include a condition in which participants learn a single set of 12 pairs in one learning period and recall these after 55min, with the rest of the procedure dropped. This would eliminate interference due to the interleaving of learning and testing, providing a valuable comparison. In addition, enhancements to the VALMT software could record the point at which each word-pair reaches criterion during the learning process, facilitating granular analysis of possible interference within a learning period.

If future work indicates that interference is, in fact, a major driver of results then providing a confound-free measure of forgetting, especially in the presence of a learning deficit and therefore differential interference, will require further development of the VALMT procedure.

A comparison with existing clinical tests highlights a trade-off in design: the VALMT equates learning and avoids differential over-learning and retrieval practice but may introduce differential interference, while many clinical tests hold the number of presentations fixed (e.g. CVLT, Delis et al., 1987; WMS-LM), which limits differential interference, but learning is not equated since material is not learnt to criterion and differential overlearning and retrieval practice can occur. The potential role of interference due to learning errors also highlights the need to record and analyse learning error rates. Many ALF studies use free-recall of a list of words as a measure of verbal memory (e.g. Butler et al, 2007; Gascoigne et al., 2012; Mameniskiene et al., 2006; Weston et al., 2018 ). In such studies the number of times the entire list is presented is

recorded and analysed as a measure of learning, but the number of individual errors made (recalling a word which was not on the list) is not. Similarly, where word-pairs are used as stimuli (e.g. Atherton et al., 2014; Mary et al., 2013) the number of times the entire list is presented is reported and analysed, but not the number of incorrect responses given. It may be that these incorrect responses made during learning are influencing the results of such studies, and by not recording and analysing these a subtle learning deficit has been missed. If that is the case then later forgetting identified as ALF in the strong sense of the term may not in fact reflect a pure retention problem.

Slower learners will also take more time in total to complete the full VALMT procedure. It is thus possible that they will experience greater fatigue, and that this may influence their results. The interleaving of learning and testing means that any increased fatigue should impact performance at all test delays. Any impact might therefore be expected to reduce scores at all delays for slow learners, rather than influencing the rate of forgetting between delays. Thus, it is hard to see how differences in fatigue could be the primary cause of the large difference in forgetting rates seen between 30min and 55min. More conclusive evidence could be provided by the previously suggested future replication including a condition in which participants learn a single set of 12 pairs in one learning period and recall these after 55min with the rest of the procedure dropped. This would greatly reduce any possibility of differential fatigue.

Finally, in the case of the VALMT, since individual pairs are only presented until they have been learnt to criterion, the pairs which are learnt faster will also experience a longer duration delay before delayed testing, compared to those which are learnt more slowly, which adds a confound, and again this would be greater for slower learners. However, in practice the maximum extra delay introduced is in the order of tens of seconds. While this might influence recall at the 5 minute delay, any impact will be negligible at the 30 and 55 minute delays.

Overall, the differences between the highlighted studies demonstrate the sensitivity of different types of paradigms for revealing subtle issues, and further demonstrate the impact of the different methodologies used (Elliot et al, 2014).

#### *2.3.4.2 Why do some participants show poorer learning performance, and what is the significance of this?*

Why do some older participants take significantly more trials to complete learning to criterion? What could be causing their slower learning?

First, it could be caused by the early preclinical stages of AD, or some other age related cognitive decline. While Experiments 1 and 2 did not include measurement of any AD biomarkers, it is known that memory complaints are a predictor of future development of MCI and AD (Mitchell et al., 2014; Weston et al., 2018). If slow learning was caused by an early stage AD process slow learners would be expected to report a higher level of memory complaints. This is indeed what was observed, with the correlation being large and significant.

Second, it may be that the Older participants were less familiar with using computers, which may make the computer based learning process harder for them in some way and thus lead to more errors. However, all participants completed a demonstration before the main study started; they were exposed to the software and could ask questions of the researcher. They were asked if they understood the task before proceeding. This would argue against familiarity with computers being the main cause of errors. Participants unfamiliar with computers may also take longer to type their answers, and/or spend more time thinking about what to do. However, the mean seconds per response during learning did not correlate with number of trials needed to learn to criterion, which also suggests familiarity with computers was not the primary cause.

Third, some older participants may have paid less attention, or may have invested less effort. While there was no direct attention measure built into the task or procedure, the fact that testing was done face to face with a researcher present, and the researcher did not observe any noticeable lack of attention, suggests that participants are likely to have paid attention to what they were doing.

Fourth, some of older participants may have found it harder to understand or follow instructions. However, the fact that there were no differences in IQ or education between the fast and slow-learning groups suggests this is unlikely.

Finally, slower learning could be directly age related, in which case older participants should make more errors. However, there was no correlation between age and learning performance in the Older group.

Overall, the evidence suggests that the slower learning some older participants displayed is caused by a cognitive deficit, in which case it is possible that VALMT learning performance provides an indicator of preclinical AD or another form of dementia. Further investigation of this possibility could include use of a standardized neuropsychological test battery and other multi-dimensional criteria (McKhann et al., 2011; Albert et al., 2011) to help minimise any possibility that an undiagnosed medical

condition has already affected memory to a level that standard clinical measures can detect, combined with longitudinal follow-up to validate the success of VALMT in predicting progression to MCI or AD.

The larger variation seen in learning performance in the Older group also reminds us that while matching individuals on the main demographic variables is the standard procedure in research where overall group performance is under investigation, in neuropsychology we should also consider individual differences since they can impact performance significantly.

#### *2.3.4.3 Is there evidence for different forgetting curves, and does forgetting start early or later?*

Whereas early theories of memory formation referred to one process of consolidation to stabilise engrams, more recent formulations have suggested a more complex process. For example, Dudai (2004) has differentiated between an early process known as 'synaptic consolidation' which occurs in the first few minutes to possibly hours in the hippocampal complex, and 'systems consolidation' which occurs over long time frames and although initially dependent on the hippocampus, becomes independent of that region over time. This has led some researchers to compare rates of forgetting at two different time points. For example, Hoefeijzers et al (2013) compared forgetting in the first 30 minutes with that between 30minutes and one week later, and found that their TEA patients were equivalent to their controls in terms of the 'early forgetting' but significantly different for the 'late forgetting'. They used this difference to argue that the ALF that the patients experienced was caused by a consolidation problem and, in particular, late consolidation. They do concede, however, that "forgetting across [the] early time interval is rather limited in both groups. This relative lack of early forgetting may be the reason we did not find a correlation between early and late forgetting rates" (Hoefeijzers et al, 2013, p1554).

Many other studies in the ALF literature which report late-onset forgetting were performed with TLE patients (e.g. Mameniskiene et al., 2006). However, in a study with TLE patients which addressed several methodological issues identified in previous work, Cassel et al. (2016) found that forgetting was detectable by 10 minutes in their visuo-spatial task, while for short stories they found forgetting developed in a more progressive manner, starting early but only reaching statistical significance after one week. Cassel et al (2016) interpreted their findings as evidence for forgetting starting during the early

consolidation stage, and suggest that differences in forgetting patterns reflect a continuum of severity and/or test sensitivity rather than a difference between early and late-onset forgetting.

While the VALMT uses word-pair learning rather than story recall, the methodology is closer in design to that of Cassel et al. (2016) in that it requires learning to criterion across all individual word pairs. A further similarity is that Experiments 1 and 2 used different word pairs for testing at different time points. Both these factors help minimise the potential problem of retrieval practice.

The Slow_Older group scored lower that the Younger and Fast_Older groups at all timepoints, even at 5min and 30min. They also showed a similar early forgetting rate (5-30mins) to the other two groups, but then an accelerated late forgetting rate (30-55mins). Taken together the lower recall at early delays combined with an accelerated forgetting seem to more strongly support an early-onset and progressively increasing forgetting pattern rather than a late-onset forgetting in this sample. However, even if ALF in TLE and in the Slow_Older group reflects early-onset forgetting, it remains possible that other participants or groups may show more pronounced forgetting at longer intervals of days or even weeks.

In a large review of forgetting rate studies, Radvansky et al. (2022) found that normal healthy forgetting does not follow a single forgetting function, as usually assumed. They found evidence of acceleration and deceleration of forgetting at different timescales. They propose a 'Memory Phases Framework' with four intervals based on cognitive and neurobiological evidence: up to 60 seconds (Working Memory), 60 seconds to 12 hours (Early Long-Term Memory, eLTM), 12 hours to 7 days (Transitional Long-Term Memory, tLTM), and beyond 7 days (Long-Lasting Memory, LLM). They found that forgetting rate increases during the WM phase, slows during eLTM, remains relatively stable during tLTM and increases again in LLM.

Experiments 1 and 2 fall within the eLTM phase, where forgetting rate should decrease as hippocampal consolidation occurs. The results show this standard pattern for the Younger and Fast_Older groups but not for the Slow_Older group, suggesting hippocampal consolidation is impaired in the Slow-Older participants. Whether there remain others within the experimental samples who show normal forgetting to 55min as measured by the VALMT but would then go on to display greater forgetting during the tLTM or LLM phases, perhaps due to problems in neocortical consolidation or retention, while theoretically plausible, remains an open question. The VALMT as currently

designed specifically targets early identification of memory deficits, within a single clinical visit. It could be adapted to monitor memory over delays of days or weeks, or alternatively the Crimes and Four Doors tests developed by Baddeley and colleagues (Baddeley et al., 2014; Baddeley et al., 2021; Laverick et al., 2021) may be more appropriate for such extended delays. Future comparisons would help clarify whether a single test can address all timeframes adequately, or whether multiple tests are required.

#### *2.3.4.4 Is the VALMT suited to early identification of those at risk of developing MCI or AD?*

Those who subjectively report suffering from memory problems can often perform normally on standardised clinical memory assessments which measure performance at delays of up to 30mins. Many explanations have been proposed for this inconsistency. One possible explanation is that subjective complaints reflect forgetting that occurs over a timeframe beyond that tested with standard clinical measures, which aligns with the late-onset forgetting conception of ALF (Manes et al., 2008; McGibbon & Jansari, 2013). Another possible explanation is that existing clinical measures are not sensitive enough to detect subtle early impairments which are the underlying cause of the complaints, which aligns better with the early-onset progressive forgetting conception of ALF (Cassel et al., 2016).

No relationship was found between complaints and recall performance for the WMS-LM tests, but a strong correlation was found with VALMT scores. Combined with the recall impairments seen with the VALMT, this provides evidence for an early-onset of forgetting. This also suggests that the VALMT (testing at delays up to 55min) is a more sensitive test than the WMS-LM (testing at the standard 30min) at detecting this and can, within the window of a single clinical visit, detect subtle impairments that are related to memory complaints, although a more conclusive comparison between the two tests will require future work testing with the WMS-LM at a 55min delay. This is important in relation to MCI and AD, as there is evidence that those who perform normally on standard tests but complain of memory problems are at greater risk of going on to develop MCI and AD (Mitchell et al., 2014; Weston et al., 2018).

There are other reasons to predict that the VALMT may be suitable for early detection of those at risk of MCI/AD. One of the first cognitive functions impacted by AD is episodic memory (Fox et al., 1998), and impairment of this sort is used as one criteria when diagnosing amnestic MCI (aMCI) which carries an increased risk of progression to

AD (Silva et al., 2012). To reliably detect such impairment at an early stage it would be advisable to use a test that relies heavily on memory and as little as possible on use of cognitive strategies. Cued-recall provides some support for recall by means of the cue-word, reducing the need for recall strategies. In addition, the use of unrelated word-pairs makes it harder to use strategies based on word categories, and avoids the scaffolding provided in story recall paradigms. Together, these elements of the VALMT reduce confounding that could be introduced by cognitive strategies and should make this test well suited to detection of early stage episodic memory impairment.

Furthermore, associative learning of the type used in the VALMT is vulnerable to the impact of early stage AD (Sapkota et al., 2017). Associative learning is also known to rely on entorhinal cortex and hippocampal brain regions which are vulnerable to change in early AD (de Rover et al., 2011; Coupe et al., 2019; Braak & Braak, 1998). So, it is to be expected that early stage AD would impact VALMT scores.

Finally, if the disease process has led to subtle learning deficits, the more challenging VALMT learning process which requires learning each pair to criterion may provide a more sensitive test than existing clinical tests such as the CVLT and WMS-LM which use a fixed number of presentations and do not require learning to criterion.

### 2.3.5 Conclusions

Experiments 1 and 2 used a variant of the VALMT word-pair learning paradigm to reveal differences in delayed recall performance between younger and healthy older participants at an earlier time point than documented previously. Importantly, those older participants who learnt more slowly also forgot more rapidly within the window of a standard clinical visit, a difference which a standardised clinical measure was unable to detect. This objective impairment of the slow learners is associated with increased subjective memory complaints.

These experiments highlight both the importance of the specific details of the methodology and the need for taking into consideration the variance between individuals that only begins to express itself as we age and which may be the underlying basis for why some people forget rapidly and others do not.

In general terms, further studies addressing these fine-grained issues are needed to elucidate the mechanisms of accelerated forgetting with ageing, as well as to support the pursuit of a clinically-reliable test for objectively capturing this forgetting. The

effectiveness of such sensitive tests in predicting future progression to MCI and AD is also worthy of future investigation.

More specifically, these results suggest four aims for future experiments within the scope of this PhD: 1) develop a procedure which does not require the interleaving of learning and testing, to investigate the possible role of interference in the observed results; 2) reduce the complexity and duration of the procedure to make a test more suited to general clinical use; 3) investigate forgetting over longer timeframes, to check whether VALMT can detect late onset rapid forgetting which may reflect a different aetiology; 4) investigate VALMT performance with participants of all ages in the general population, to investigate when the deterioration seen in Older groups starts, and how performance develops across the total lifespan.

# 3 Chapter 3 – Development and use of an online version of VALMT

## 3.1 Introduction

The studies reported in Chapter 2 identified several important results worthy of further investigation and validation. Using VALMT it was possible to identify differences in the memory performance of Younger and Older groups by 55 minutes. On closer investigation, these group differences were driven largely by slower learners within the Older group, who require more trials to reach criterion and display more rapid forgetting. The performance of this Older_Slow group also correlated with subjective memory complaints, a known precursor to dementia. These results raise the interesting possibility that VALMT may, within the window of a single clinical visit, be able to identify those at risk of developing dementia in future. In contrast to the VALMT, the WMS-LM was unable to identify any differences in group performance, suggesting that VALMT may be better able to identify subtle memory deficits. However, the results also highlight some concerns and opportunities for further development of VALMT, some of which link directly to the research questions prioritised in Chapter 1.

First, the procedure used in Experiments 1 and 2 was complex, with many interleaved learning and test stages. While this was done on purpose to equalise the impact of task learning, fatigue and stress when learning and testing multiple lists, it is possible that this interleaving may introduce interference and that some groups may be more sensitive to interference leading to differences in delayed recall. It is also possible that this interference may be greater for the Older_Slow group who take more trials to complete learning, and that this may then drive this groups greater forgetting. There is some evidence that healthy older participants are, indeed, more sensitive to interference in some memory paradigms; however, it is not clear how directly this relates to the specific paradigm used in this experiment. For example, inhibitory deficit theory (Hasher & Zacks, 1988) proposes that older individuals are less able to control the contents of working memory, leading to increased impact of irrelevant information, which in turn leads to poorer long-term memory performance. This is typically tested by introducing distractor stimuli or tasks during the encoding phase. For example, Mund et al. (2012) found that introducing distractor words into short story texts caused a greater decrease in subsequent delayed free recall in older participants. However studies of this type do not equalise learning, so it unclear what the impact would be where a learning criterion is

enforced. Davis et al. (2003) adopted a proactive interference approach to investigating sensitivity to interference. Their participants first learnt a list of 15 words across 5 trials, then after testing delayed recall at 24hr they learnt a second list of 15 new words, using only one trial. The number recalled from this second list was compared with the number recalled on the first trial of the first list. They found no effect of age, indicating that older participants were not more sensitive to interference. Since the existing evidence is mixed, it is desirable to eliminate these possible interference confounds by developing a procedure which does not require such complex interleaving, and check whether the same pattern of results is seen.

Second, the complex procedure also requires careful preparation and administration by the researcher, and requires face-to-face testing. Together, these factors make it unsuitable for large scale testing, as participation is limited by both the availability of trained researchers and the need for travel by either the participant or the researcher. The face-to-face requirement also makes it unsuitable for repeated testing or testing at longer delays, as that would require multiple meetings between researcher and participant. It is therefore desirable to develop a test that can be administered remotely, and preferably in a fully automated manner. The Covid-19 pandemic and the consequent restrictions on travel and face-to-face meetings further highlighted a general need for tests of this type, which are suited to a tele-assessment approach.

Third, the longest delay in the previous studies was 55mins. This was selected based on previous work with VALMT in epilepsy (McGibbon & Jansari, 2013) and the desire to look for differences that may be detectable in a single clinical visit. However, this means any forgetting which starts at a longer delay of several hours or days will not have been captured. It is desirable to check whether VALMT can detect such late onset rapid forgetting which may reflect a different aetiology.

Finally, the focus of the studies in Chapter 2 was on the difference between younger and older participants. While this was designed to effectively check for any impact of healthy ageing with minimal sample sizes, it leaves open the question of when deterioration of memory performance begins, and how performance develops across the total lifespan. To clarify this, it would be beneficial to test with participants of all ages in the general population.

To deal with the points raised above, a fully automated online version of VALMT was developed. All steps of the procedure were automated, including recording informed consent, learning, delayed test and post-participation debrief. By using an online version

the test can be made available to anyone globally; the only requirement is to have access to a web browser and an internet connection. By automating the entire process there is no need for a trained researcher to be involved in data collection. An online test can easily scale to thousands of participants per study, and is well suited to repeated testing and testing at longer delays as no travel for face-to-face testing is required. This new online version was designed to replicate the face-to-face learning and test procedure as closely possible, with two key exceptions.

First, to simplify the procedure only one list of 12 word-pairs is learnt, requiring just one learning period. This single set of pairs is then tested at 55mins, with no other learning or testing performed during the delay. This greatly reduces any potential for confounds due to differential interference or differences in sensitivity to interference. This simplified procedure is also better suited to an online large scale test with the general public or in a clinical setting where long complex learning procedures are not practical. The 55min delay was selected because the most important group differences in previous studies became significant at this delay, and it was important to check whether the online test found the same differences.

Second, the same set of 12 pairs were tested again at a second longer delay, to look for forgetting which starts beyond 55mins. While it would be better to test unique pairs at each delay, this is not possible when only one set of pairs can be learnt. In this case, a repeated recall must be used. The first recall at 55min can be expected to consolidate memory and reduce future forgetting, so the second test must be at a long enough delay to ensure some additional forgetting is still expected. Based on previous studies 24 hours was selected as the second delay. This has the advantage of including a night sleep, so should detect deficits due to problems with consolidation during sleep.

Additional statistics were built into the new version to facilitate more granular analyses of results. In particular, whereas the face-to-face version recorded only the total number of trials taken to complete learning, the online version records the result of every trial. This allows the relative difficulty of each pair to be compared, and allows the impact of learning errors to be evaluated at a granular per-pair level.

The online VALMT demographics questions included administration of the memory complaints questionnaire used in Experiment 2; however the wording of two of the questions was adjusted to suit the automated multiple choice answer format. The PSQI (sleep questionnaire) was not included as Experiment 2 showed this provided little value.

The following experiments are reported, evaluating this new online VALMT and using it to investigate forgetting within a range of populations:

- Experiment 3: Initial validation using a sample of first year undergraduate students. This aimed to verify that the online version works as expected and provides similar results to the face-to-face testing of younger participants, and to investigate memory performance over the longer 55min to 24hr interval in a young age group.

- Experiment 4: Testing with the general population, across all ages from 18 upwards. This aimed to investigate: how scores and age-related group differences compare to those seen with the face-to-face version; the impact of reducing the possible confounds due to interference; the relationship between errors made during learning and the subsequent delayed recall scores; memory performance over the longer 55min to 24hr interval in older age groups; how learning and memory performance change across the lifespan in healthy ageing.

- Experiment 5: Testing with participants in the 16-17yr age range, recruited from UK schools. This is the first study of its type, looking at memory and forgetting over extended delays in this younger group. This is one of a very few studies of its type, looking at memory and forgetting over extended delays in this younger group.

- Experiment 6: Testing with a group for whom face recognition memory data is available, to investigate whether group differences in memory performance detected by VALMT are specific to the verbal domain or correlate with a non-verbal form of memory.

## 3.2 Experiment 3 - First validation of the online VALMT with younger participants

### 3.2.1 Rationale

This experiment had three primary aims. First, to test the online VALMT procedure, and highlight any issues with the software or process. Second, to compare results for the new online VALMT with those from the previous face-to-face version for young healthy participants as a check of concurrent validity. Third, to investigate memory performance over the longer 55min to 24hr interval in a young age group.

### 3.2.2 Methods

#### *3.2.2.1 Participants*

Participants were recruited by advertising the study on Goldsmiths College's Research Participation Scheme (RPS). First year undergraduates were recruited, and took part for research credits. For the comparisons with results from face-to-face testing (N=30) power analysis indicated a minimum sample size of 15 was required to detect at least a large effect (Cohen's d = 0.80), and a minimum of 145 to detect at least a medium effect (d = 0.50), assuming a desired power of 80%.

##### 3.2.2.1.1 Inclusion criteria

To be included in the analyses a participant had to meet the following requirements, where the number in [brackets] is the number who failed to meet that criteria. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as many excluded participants failed multiple criteria:

1. Must not report dyslexia [6]
2. Must not report a medical condition that might impact memory [1]
3. Must be aged under 30yrs [3]
4. Reported English language level must be either first language or fluent [3]
5. Must learn all 12 pairs to criterion within 20 minutes [8]
6. Must complete all 3 stages: learning, 55min & 24hr tests [10]

Participants were instructed to complete the delayed tests as close to the requested time as possible (55mins, 24hrs), but to still complete the delayed tests even if they were unable to come back at the exact time. The intention was that data from those returning at the correct time could be used to compare results with previous face-to-face studies, while data from those returning early or late could be used to explore the impact of variation in the test delays. The additional inclusion criteria for the 'All_Criteria_Met' group to be compared across studies were:

7. Must complete the 55min test between 45 and 65min (55min +/- 10min) [21]
8. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [15]

There is a potential for exclusions to distort the results if there is any relationship between participants' memory and their failing to complete the later tests at the correct delay. For example, perhaps those with a higher rate of forgetting are more likely to forget to return for later tests, or forget to return at the correct time. To analyse this risk a group comparison was performed between those who met all criteria (N=49) and those who met criteria 1,2,3,4 & 5 but failed to meet any of criteria 6 , 7 & 8 (either failed to complete the delayed tests, or did these outside the acceptable time window; N=35). The two groups were compared on the number of errors made when learning to criterion. This variable is the best proxy available for memory performance in the excluded group, since recall scores cannot be used as many did not complete the delayed recall stages. This comparison showed that the two groups were not significantly different on errors made, which suggests excluding these participants will not significantly impact the results ($Errors_{included}$ = 20.00, $Errors_{excluded}$ = 24.00, MWU = 736, p=.27, r = 0.14, BF = 0.57).

#### 3.2.2.1.2 Included participants

A total of 107 individuals took part. After applying all exclusion criteria (1-8) 49 participants were included in the cross-study All_Criteria_Met group. A further 20 met criteria 1 to 6, but not criteria 7 and/or 8 (completed all 3 stages, but delayed tests not taken at the correct delays), and were included in a larger 'Delay_variation' group to explore the impact of variation in test delays. The demographics of these groups are summarised in Table 3.1. As some results will be compared with the Younger group from face-to-face testing in Experiment 2, the demographics for that group are also shown.

*Table 3.1 Demographic information as a function of group*

| Variable | All_Criteria_Met group (N=49) | Delay_variation group (N=69) | Experiment 2 Younger (N=30) |
|---|---|---|---|
| Gender | 11M/38F | 13M/56F | 6M/21F |
| Age Mean(SD) | 19.3 (1.85) | 19.2 (2.78) | 24.83 (2.87) |
| Education: | | | |
| Batchelors | 0 | 1 | 11 |
| Diploma | 3 | 4 | n/a |
| Doctorate | 0 | 0 | n/a |
| High School | 45 | 63 | 9 |
| Masters | 0 | 0 | 7 |
| Secondary | 1 | 1 | 3 |
| Technical | 0 | 0 | n/a |
| Language: | | | |
| First Language | 33 | 50 | 17 |
| Fluent | 16 | 19 | 13 |
| Total MCS | 2.33(2.34) | 2.78(2.62) | 1.83(1.64)* |

*Memory Complaints Scale (MCS) total for Expt 2 has been adjusted to match the coding scheme used in the online VALMT

The All_Criteria_Met group and the Younger face-to-face group from Experiment 2 were matched on Gender ($X^2(1) = 0.56$, $p = .45$, BF = 0.40), Language($X^2(1) = 0.91$, $p = .34$, BF = 0.43) and total MCS score (t(77) = 1.01, p = 0.32, Cohen's d = 0.23, BF = 0.37), but were significantly different on Age, being on average 5.5 years older ($Age_{Online} = 19.3$, $Age_{Face\text{-}to\text{-}face} = 24.83$, t(77) = 10.38, p < .001, Cohen's d = 2.4, BF = $1.45*10^{13}$). Although this age difference is statistically significant, it is not expected to indicate a large difference in cognitive function. It is clear that their Education levels are not matched, as 60% of the face-to-face group have completed a Batchelors or Masters degree, while none of the All_Criteria_Met group have. This reflects the younger age of the All_Criteria_Met group and the fact they are first year undergraduates. However, it is expected that the majority will go on to complete their Batchelors degree, at which point the groups would be closely matched, so this difference in current Education levels is not expected to indicate a difference in level of IQ or other cognitive function. Overall, although the groups differ on Education and Age, cognitive function should be well enough matched to make comparisons of face-to-face and online VALMT valid.

#### *3.2.2.2 Stimuli*

The stimuli set consisted of a subset of the words in the original McGibbon and Jansari (2013) study, which were matched for familiarity, concreteness, imageability and frequency. All words were nouns, two syllables and 4-5 letters long. Words were randomly assigned to pairs to create 12 unrelated pairs (refer to Appendix B for detail). Words in any pairs with obvious semantic relationships were re-paired.

#### *3.2.2.3 Procedure*

The entire VALMT procedure, including gathering of demographics, recording consent, learning and testing was performed online. All interaction was through the participants' standard web-browser; no software or application had to be installed. This meant participants could use any web connected device (PC, tablet, smartphone etc.). Text was displayed in black characters on a white background in the browser's medium font size. Participants could participate in their own time, whenever convenient. They were asked to do this when they were in a quiet location and would not be disturbed.

Learning was performed to criterion in an identical manner to the face-to-face version (refer to Experiment 1 for detail).

At delayed test the participants were presented with the first word of a pair (e.g. TROOP-???) and were required to type in the second word of the pair. No feedback was provided until the end of the test, when the total number correct was presented. Testing of all 12 pairs was conducted at a delay of 55min, and repeated at a delay of 24hr.

### 3.2.3 Results

#### *3.2.3.1 Learning performance*

To compare the online and face-to-face VALMT versions the number of trials needed to learn to criterion in this experiment was compared with the equivalent number for the Younger group from Experiment 2. In the reported face-to-face studies (Experiments 1 & 2) participants learnt three sets of 12 pairs in three separate learning periods and it is possible their learning performance may have improved as they got more familiar with the task. Using the trials from the first learning period provides the best comparison with the online version, in which there is only one learning period. For this reason Table 3.2 reports the number of trials taken during the first learning period of Experiment 2, rather than the mean across all 3 learning periods.

*Table 3.2 Trials to learn to criterion for online and face-to-face VALMT*

| Factor | Online VALMT All_Criteria_Met group (N=49) *Mean(SD)* | Face-to-face VALMT Experiment 2 Younger group (N=30) *Mean(SD)* |
|---|---|---|
| Trials | 57.96(17.98) | 52.43(14.03) |

There was no significant difference between the trials to learn to criterion for the online and face-to-face versions ($Mdn_{online}$ = 56.00, $Mdn_{face\text{-}to\text{-}face}$ = 51.00; *MWU* = 854, *p* = .23, *r* = 0.16, BF = 0.32), suggesting the two versions operate in a similar manner.

### *3.2.3.2 Delayed cued-recall performance*

To investigate delayed recall performance and forgetting the mean recall scores for the All_Criteria_Met group at 55min and 24hr are plotted in Figure 3.1.

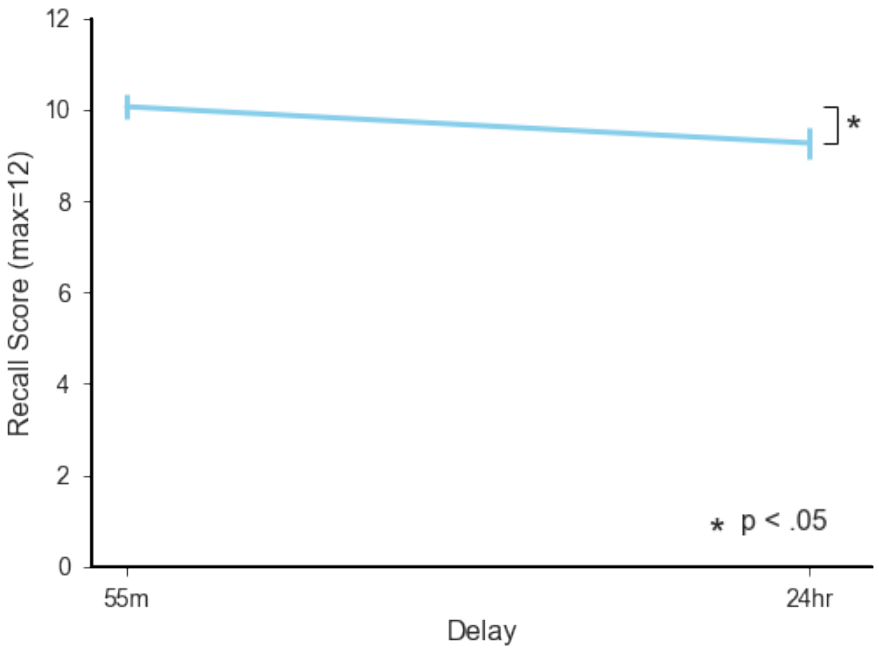


*Figure 3.1 Delayed recall performance at 55min and 24hr delays for the online VALMT All_Criteria_Met group (N=49, error bars +/- 1SE).*

The group showed a small but statistically significant amount of forgetting between 55m and 24hr ($Recall_{55m}$ = 10.06, $Recall_{24hrs}$ = 9.27, t(48) = 3.31, p=.002, Cohen's d = 0.37, BF = 17.2).

To compare the online and face-to-face VALMT, Table 3.3 shows the mean cued recall score at 55min (the only common delay) for the All_Criteria_Met group using the new online version and the Younger group from the face-to-face version (Experiment 2).

*Table 3.3 Cued recall scores at 55min delay for online and face-to-face VALMT*

| Factor | Online VALMT All_Criteria_Met group (N=49) *Mean(SD)* | Face-to-face VALMT Experiment 2 Younger group (N=30) *Mean(SD)* |
|---|---|---|
| Trials | 10.06(1.80) | 10.03(1.89) |

There was no significant difference between the 55min cued recall scores for the online and face-to-face versions ($M_{online}$ = 10.06, $M_{face\text{-}to\text{-}face}$ = 10.03; $t(77) = 0.06$, $p = .95$, $d = 0.02$, BF = 0.24), again suggesting the two versions operate in a similar manner.

#### 3.2.3.2.1 Impact of variation in test delay

The inclusion criteria for the main analyses of online VALMT data include acceptable time windows for completion of the 55m and 24hr delayed tests. The window sizes are +/- 10mins for the 55min delay and +/- 4hrs for the 24hr delay. To validate these criteria the relationship between test delay and cued-recall score was investigated. To provide the widest possible range of delays this analysis was performed for the Delay_variation group, which includes those who completed their delayed tests outside the acceptable windows. This may also provide some indication of the likely impact of extending the 24hr delay to a longer interval. The scatterplots in Figure 3.2 illustrate the relationships for this group.

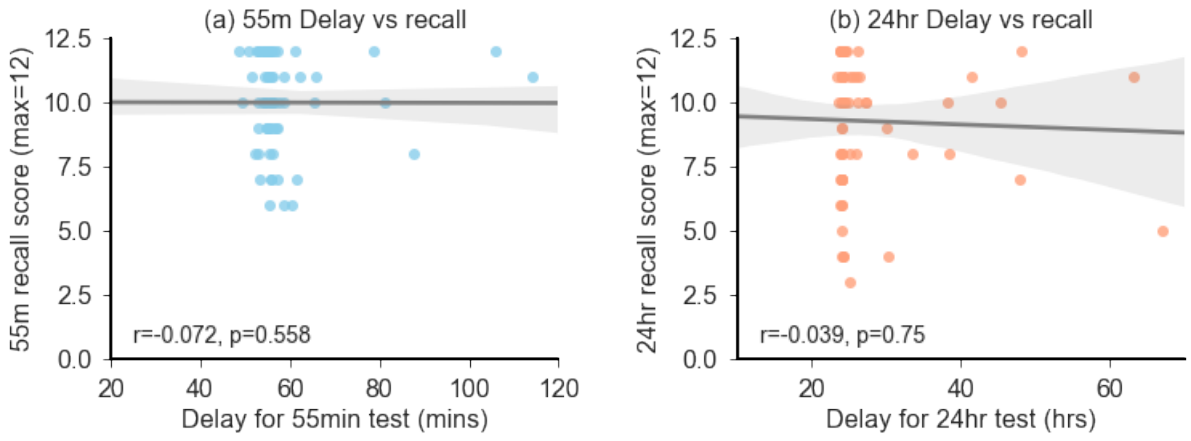


*Figure 3.2 Correlation of cued-recall scores with test delay, for (a) the 55m and (b) the 24hr delayed tests, for the Delay_variation group (N=69); shaded area is 95% confidence interval.*

There was negligible correlation between test delay and recall score for either the 55min (r = .07, p = .56, BF = 0.18) or 24hr (r = 0.04, p = .75, BF = 0.16) tests (see Figures 3.2a & 3.2b respectively). This suggests variation in the timing of the delayed tests within the acceptable windows is unlikely to have impacted recall scores, and the windows are therefore acceptable. There were only two participants who completed their 24hr test at beyond 2 days, so there is no clear evidence for the likely impact of extending the second test delay to longer delays such as three days or one week.

#### *3.2.3.3 Distribution of learning errors across word-pairs and their relationship to recall*

To evaluate the effectiveness of each word-pair, and the relationship between learning errors and subsequent recall, Figure 3.3 illustrates the mean number of errors made for each individual pair, and the corresponding mean recall rate at each delay, for the All_Criteria_Met group.

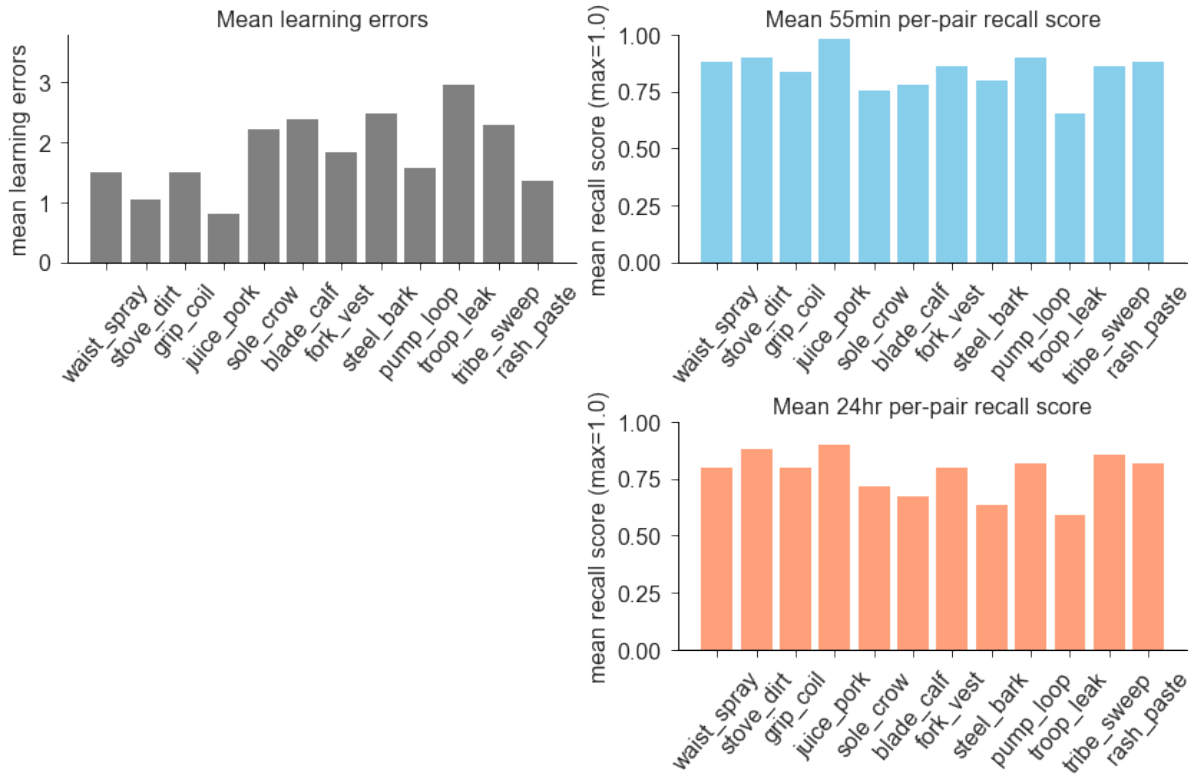


*Figure 3.3 Mean learning errors and delayed recall by individual word-pair for the All_Criteria_Met group (N=49).*

The data shows variation in difficulty between pairs which is beneficial for distinguishing between high and low performers. No pairs encounter zero or a very high

number of errors, or zero or 100% recall, so all pairs are providing useful information. Visual inspection shows that those pairs which result in most errors are recalled more poorly at both delays, while those pairs which generate fewest errors are recalled best.

#### *3.2.3.4 Relationship between delayed recall and learning errors*

To further investigate the relationship between errors made during learning and subsequent delayed recall, the scatterplots in Figure 3.4 illustrate the association between the total number of errors made and the cued recall scores at each delay for the All_Criteria_Met group.

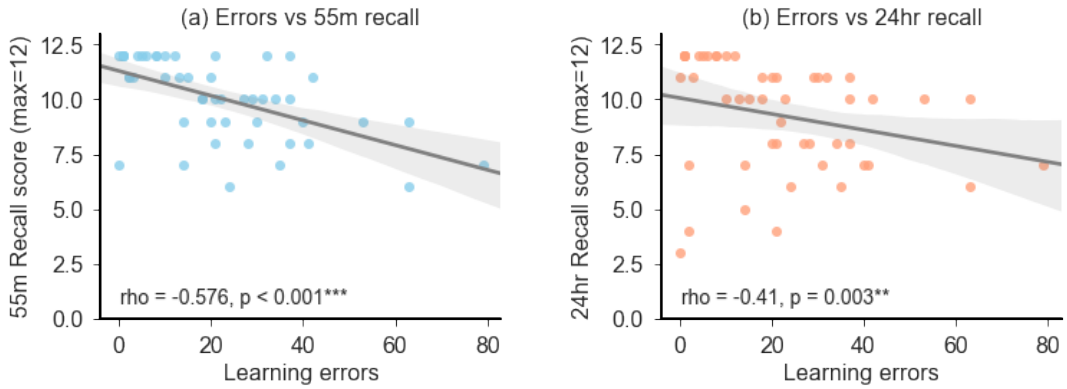


*Figure 3.4 Correlation of total learning errors with delayed recall at (a) 55m and (b) 24hr for the All_Criteria_Met group (N=49); shaded area is 95% confidence interval.*

There was a significant negative correlation between learning errors and delayed recall at both 55min and 24hr (55min rho = -0.58, $p < .001$; 24hr rho = -0.41, $p = .003$; see Figures 3.4a & 3.4b respectively). This indicates a strong relationship between these variables, with those who make the most errors subsequently recalling the fewest pairs.

To investigate this at a per-pair granularity the scatterplots in Figure 3.5 illustrate the relationship between the mean number of errors made by the All_Criteria_Met group members when learning each individual pair and the subsequent mean recall scores for that specific pair. These plots show one data-point for each of the 12 word-pairs.

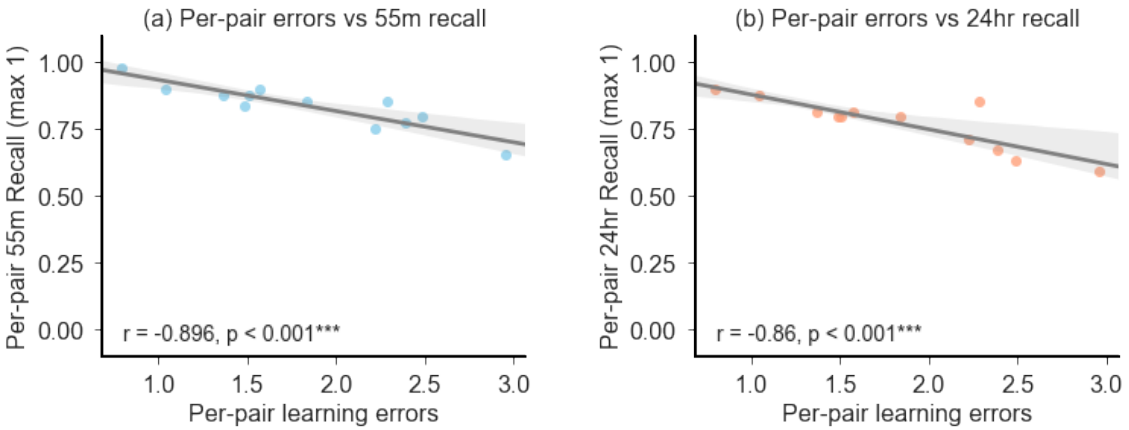


*Figure 3.5 Correlation of learning errors per word-pair with delayed recall per word-pair at (a) 55m and (b) 24hr for the All_Criteria_Met group (N=49); shaded area is 95% confidence interval.*

There was a very large and significant negative correlation between the per-pair learning errors and delayed recall at both 55min and 24hr for the All_Criteria_Met group (55min r = -0.896, p < .001, BF = 331; 24hr r = -0.86, p < .001, BF = 103; see Figures 3.5a & 3.5b respectively). This confirms the visual pattern seen in Figure 3.3, in which those pairs which encounter most errors are recalled most poorly.

### 3.2.4 Discussion

#### *3.2.4.1 Online testing procedure*

This experiment was the first to use a new online version of the VALMT. Over 100 participants took part, and none reported any issues with the online process or technical issues with the test, despite accessing it from a range of devices and operating systems. This shows that the VALMT can operate successfully in an online manner.

There was, however, a greater attrition rate than with the face-to-face test. Many people dropped out before completing all stages, or completed a delayed test outside the normal acceptable window. In the case of face-to-face testing, the researcher is in control of the timing all aspects of the process and the face-to-face nature makes it less likely a participant will drop out. Since unsupervised online testing lacks these characteristics a greater attrition is to be expected. This has the potential to distort results if there is any association between dropping out and memory performance, for example if those with poorer memory are more likely to drop out. However, analysis of learning performance for those who dropped out showed no significant difference compared to those who completed all stages. This suggests the dropouts are unlikely to have distorted results for

this younger sample. To help reduce attrition there may be value in adding automated reminders in future, particularly for the 24hr test. However, this has implications for data privacy regulations (e.g. European Union's GDPR) and is beyond the scope of this work.

A further factor influencing inclusion rate is the size of the acceptable time windows for starting the delayed tests. The wider the windows, the easier it is for participants to complete their delayed tests at an acceptable time, and the fewer participants that will be excluded for being either late or early. For the 55min test a window of +/- 10mins was used, while for the 24hr test the window was +/- 4hrs. To validate the size of these windows it was important to know if the delayed recall scores vary significantly across the window. For example, do participants completing their test at the shortest acceptable delay score higher than people completing their test at the longest acceptable delay? Analyses showed that there was negligible correlation of test delay with recall score for either delay, so there is no need to use tighter windows.

#### *3.2.4.2 Comparison with previous results*

It is desirable that the new online VALMT provide similar results to the face-to-face version used in previous studies, as this would provide a level of concurrent validity, suggesting both versions are testing the same thing. The two tests can be directly compared on two variables: trials required to learn to criterion, and cued-recall at the 55min delay. Comparison with the results from the Younger group in Experiment 2 (face-to-face testing) showed no significant difference on either variable, suggesting the two versions do indeed provide similar results.

#### *3.2.4.3 Detailed analysis of learning errors and implications for the role of interference*

The analyses of Experiment 1 and 2 in Chapter 2 identified possible confounds due to differential interference or differences in sensitivity to interference. The procedure for the online version deliberately eliminated the complex interleaving of learning and testing present in these previous experiments, reducing the opportunity for interference to play a role. The fact that the 55min delayed recall scores are similar for Experiment 2 (with the interleaved procedure) and the online version (no interleaving) suggests that interference due to interleaving is unlikely to have influenced results for the Younger groups in the earlier experiments. Equivalent comparisons with Older groups will be performed in later experiments.

However, even with the interleaving of learning and testing eliminated there can still be interference due to errors made when learning to criterion during the single learning session. Each pair must be recalled correctly three times to complete learning. Once a pair has been learnt to criterion it stops being represented, while testing and presentation of the remaining pairs continues. This may create interference towards the previously learnt pairs. In addition, the errors made during unsuccessful recalls of a given pair may create interference for that pair. Indeed, as explained in the previous chapter, there is some evidence that using procedures that eliminate errors during learning ('errorless learning') can be beneficial for healthy elderly (Baddeley & Wilson, 1994; Wilson et al., 1994) and effective as a memory rehabilitation technique for AD patients (Clare et al., 2002). The additional statistics captured by the online version allow this to be investigated in detail, making it possible to look at both the total errors made during learning, and the errors made for each individual word-pair.

For the current experiment the total errors made correlated significantly with the delayed recall scores at both delays. However, the correlation between errors made for each individual word-pair and subsequent recall of that pair is even stronger. This suggests that the key factor is not the total number of errors made, but rather where those errors are made; the pairs which generate the most errors are the ones which are recalled most poorly. One possible causal explanation for this relationship is that errors during learning cause interference which then impacts later recall. However, it could also be that word-pairs which are harder to learn are also most easily forgotten, perhaps due to difficulty in linking the two words to each other and to existing knowledge, and so greater word-pair difficulty causes both higher errors rates and lower delayed recall.

The analysis of per-pair errors also shows that the individual pairings selected perform well. They display a range of difficulties which will help to distinguish between high and low performers. Having some more difficult pairs allows differences between high performers to be identified, while having some easier pairs helps identify differences between lower performers. The fact that no pairs were recalled or forgotten by everyone shows each pair is adding value. Together, these results suggests the existing pairs can be retained; there is no need to replace any words or pairs.

#### *3.2.4.4 Forgetting between 55min and 24hr*

This is the first experiment to use VALMT to investigate memory and forgetting at the longer 24hr delay. This delay was chosen to test memory at a longer delay than previous VALMT work, with the aim of identifying any late-onset forgetting. The same pairs are recalled at both 55min and 24hr. While it is likely that the recall at 55min may provide some recall support through retrieval practice, it was thought that 24hr would be long enough to allow further forgetting to manifest.

Comparison of recall across delays found that the young participants in this experiment showed a relatively small amount of forgetting between 55min and 24hr (Cohen's d = 0.37, a small effect by convention), although this was statistically significant. Any decision on whether the second delay should be extended, for example to three days or a week, will need to wait until testing has been performed with older participants.

### 3.2.5 Conclusion

This experiment showed that a new online version of VALMT operates successfully, producing similar results to the older face-to-face version. The online nature did lead to a higher attrition rate, and although this does not appear to influence the results, further work to reduce this by adding automated reminders would be beneficial.

The online version eliminates the complex interleaving of learning and testing present in the face-to-face experiments, and the fact that group means are very similar to those from face-to-face testing suggests the interleaving is unlikely to have influenced the Younger group in previous experiments. Instead, it is the errors made during learning that appear to be the biggest driver of delayed recall scores, with the pairs that generate the most errors being recalled most poorly.

There was relatively little forgetting between 55min and 24hr, but any decisions about adjusting or eliminating the extended 24hr delay will need to wait until testing with older groups has been completed.

## 3.3 Experiment 4 - Online VALMT testing with Younger and Older groups

### 3.3.1 Rationale

Experiment 3 with first year undergraduates indicated that the new online system works as expected for younger participants, and provides similar results to the face-to-face version of VALMT for this age group. Experiment 4 was intended to extend the validation of the online version to participants across the full adult age range and who were recruited from the general population rather than from a specific university population, and also to perform additional detailed analysis made possible by the extra data captured by the new VALMT version. There were five main aims.

First, by including older participants, verify that the new version works as expected for an older population and that testing using this new version replicates the key results from the face-to-face testing of younger and older groups performed in Experiments 1 and 2. If the results are replicated this will provide evidence that the face-to-face and online versions are measuring the same thing, and provide further support for the previous findings from Experiments 1 and 2. These key previous findings were that by 55min an Older group (aged 60yrs or more) scored lower on delayed recall than a Younger group (aged 18-30yrs), that this difference was largely driven by a subset of Older participants who take longer to complete learning, and that being an Older slow learner correlated with increased subjective memory complaints.

Second, by using the new online version which eliminates the complex interleaving of learning and testing used in the face-to-face version, investigate the impact of reducing the possible confounds due to differential interference and differences in sensitivity to interference that this interleaving introduced, and thereby help evaluate whether interference is likely to have been a key driver of the group differences seen in Experiments 1 and 2.

Third, use of the granular statistics gathered by the new online system to investigate in depth the relationship between errors made during learning and the subsequent delayed recall scores. This analysis should attempt to tease apart whether the poor delayed recall displayed by some older participants is due to interference caused by errors made during learning, or reflects a greater underlying forgetting rate (perhaps reflecting a trace decay process), or is caused by a combination of these factors.

Fourth, extend the investigation of memory performance over the longer 55min to 24hr interval performed with young participants in Experiment 3 by repeating this with older participants. In particular, look for any evidence of ALF in the Older group.

Fifth, by sampling across the entire adult age range, rather than having only a younger and older group, investigate how learning and memory performance at extended delays vary across the lifespan in healthy ageing, including when any deterioration starts, thus addressing one of the identified research questions for this PhD.

### 3.3.2 Methods

#### *3.3.2.1 Participants*

##### 3.3.2.1.1 Recruitment

Participants were recruited by advertising on social media. The recruitment targeted all ages and memory abilities, stating the study was open to anyone aged 18yrs or over with a good standard of English. Based on effect sizes seen in previous VALMT studies (Cohen's d approx 1.0), statistical power analysis indicated the Younger (<30yrs) vs Older (>60yrs) group comparisons required sample sizes of at least 17 in each group to provide power of greater than 80%.

##### 3.3.2.1.2 Inclusion criteria

To be included in the main analyses a participant had to meet the following requirements, where the number in [brackets] is the number who failed to meet that criteria. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as many excluded participants people failed multiple criteria:

1. Must not report dyslexia [11]
2. Must not report a medical condition that might impact memory [21]
3. Reported English language level must be either first language or fluent [2]
4. Must learn all 12 pairs to criterion within 20 minutes [25]
5. Must complete all 3 stages: learning, 55min & 24hr tests [70]
6. Must complete the 55min test between 45 and 65min (55min +/- 10min) [41]
7. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [9]

As with Experiment 3, the inclusion criteria were used to prepare two groups of participants. First, a 'All_Criteria_Met' group (N=104) who met all criteria, which was

used for the main analyses including splitting into age subgroups and for cross-experiment comparisons. Second, a larger 'Delay_variation' group (N= 154) containing those who met criteria 1 to 5, but completed one or more of their delayed tests outside the acceptable windows. This group was used to analyse the impact of variation in test delay.

The risk of distortion of results due to excluding those who dropped out or failed to complete delayed tests at the correct time was analysed. A group comparison was performed between those who met all criteria and those who met criteria 1, 2 , 3 and 4 but failed to meet any of the criteria 5, 6 or 7. The two groups were compared on the number of errors made during learning. This variable is the best proxy available for memory performance in the excluded group. This analysis showed that the two groups were not significantly different on errors made during learning, which suggests excluding these participants will not impact the results ($Errors_{excluded}$ = 11.00, $Errors_{included}$ = 13.00, MWU = 5919, p=.33, r = .08, BF = 0.20).

#### 3.3.2.1.3 Included participants

To allow group comparisons of younger vs older participants two subsets were created from the All_Criteria_Met group: a Younger group aged 18-30yrs, and an Older group aged 60yrs and over. These specific age limits were chosen to match previous studies and thus facilitate direct cross-study comparisons. The demographics of the groups and age based subgroups are summarised in Table 3.4.

*Table 3.4 Demographic information as a function of group*

| Factor | All_Criteria_Met | Delay_variation | Younger | Older |
|---|---|---|---|---|
| N | 104 | 154 | 20 | 32 |
| Gender | 25M/79F | 38M/116F | 4M/16F | 10M/22F |
| Age Mean(SD) | 48.8 (17.6) | 46.9 (17.8) | 21.3 (3.7) | 68.25 (5.1) |
| Education: | | | | |
| Batchelors | 37 | 49 | 6 | 10 |
| Diploma | 8 | 11 | 0 | 4 |
| Doctorate | 5 | 7 | 0 | 0 |
| High School | 18 | 27 | 9 | 5 |
| Masters | 26 | 42 | 3 | 6 |
| Secondary | 7 | 13 | 2 | 4 |
| Technical | 3 | 5 | 0 | 3 |
| Language: | | | | |
| First Language | 95 | 138 | 15 | 29 |
| Fluent | 9 | 16 | 5 | 3 |

The Younger and Older groups were matched on Education ($X^2(6) = 8.49$, $p = .13$, BF = 0.93), Language ($X^2(1) = 2.31$, $p = .13$, BF = 1.28) and Gender ($X^2(1) = 0.79$, $p = .37$, BF = 0.51).

The age distribution for the All_Criteria_Met group is illustrated in Figure 3.6.

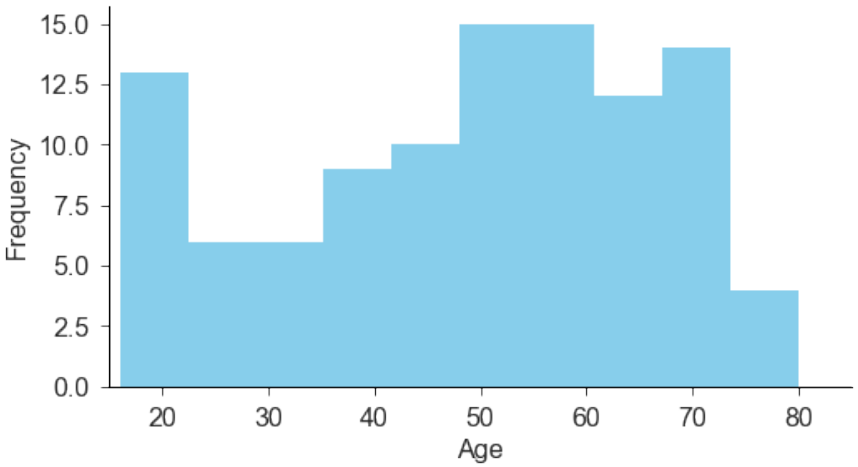


*Figure 3.6 Age distribution for the All_Criteria_Met group (N=104).*

This shows that advertising was successful in recruiting a wide range of ages, although with a relatively lower proportion in the 25 to 50 years age range.

#### *3.3.2.2 Stimuli*

The stimuli set consisted of the same 12 unrelated word-pairs used in the previous experiment. Refer to Experiment 3 for detail.

#### *3.3.2.3 Procedure*

The entire procedure, including gathering of demographics, providing consent, learning and testing was performed online using the new online VALMT. Refer to Experiment 3 for details.

### 3.3.3 Results

#### *3.3.3.1 General note on choice of groups for correlation analyses*

This experiment is the first to recruit participants from all ages across the lifespan, and analyses of correlations reported below include results for the All_Criteria_Met group which includes all these participants. Many correlation analyses are also reported separately for the Older group. This is to help investigate the consequences and causes of

a key result from Experiments 1 & 2; specifically that within the Older group there were some who learn more slowly, taking more trials to reach criterion, and that these participants displayed increased forgetting compared to their faster learning peers.

#### *3.3.3.2 Learning performance:*

To investigate variation in learning performance Figure 3.7 shows the distribution of trials to criterion for each group. The shape of the distribution is similar for all groups, with no strong indication of the bimodal split in the Older group seen in Experiment 1 (see section A.3).

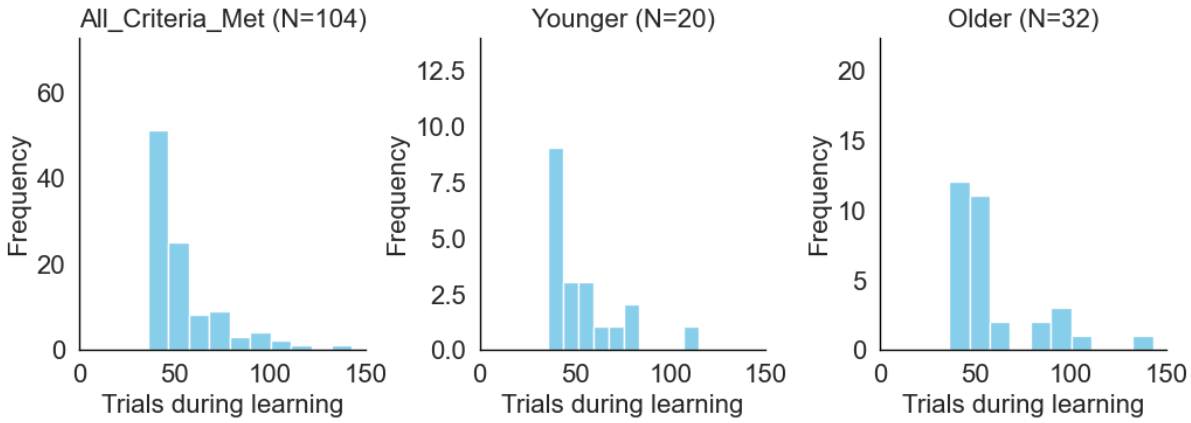


*Figure 3.7 Distribution of trials required to reach criterion during learning, as a function of group.*

To investigate the impact of healthy ageing an independent samples t-test was used to compare the Younger and Older groups on number of trials needed to reach criterion. In common with previous face-to-face studies (Experiments 1 & 2) the Older group took more trials to reach criterion, and showed greater variance. However, unlike previous studies neither of these differences reached significance ($Mdn_{Older}$ = 48.0 trials, $Mdn_{Younger}$ = 44.5 trials, MWU = 370, p = .35, r = 0.16, BF = 0.44; Older $\sigma^2$ = 648.8, Younger $\sigma^2$ = 401.0, Levene's test for equality of variances not violated, $p$ = .35).

Following the process used in previous studies (Experiments 1 & 2), to facilitate investigation of the relationship between learning performance and other variables the Older group was divided using a median split into those who required fewer trials (Older_Fast, $n$ = 13) and those who required more trials (Older_Slow, $n$ = 19). The split point was set at the median value, 48 trials.

While there was no significant difference between the Fast_Older and Younger groups in number of trials to reach criterion ($Mdn_{Fast\_Older}$ = 39.0 trials, $Mdn_{Younger}$ = 44.5 trials; $MWU$ = 84.5, $p$ = .10, $r$ = .35, $BF_{10}$ = 0.76), the Slow_Older group took

significantly more trials than the Younger group ($Mdn_{Slow\_Older}$ = 54.0 trials, $Mdn_{Younger}$ = 44.5 trials; $MWU$ = 285, Bonferroni adjusted $p$ = .016, $r$ = .50, $BF_{10}$ = 6.35).

A comparison of the two Older groups showed no significant difference in mean age ($M_{Older\_Fast}$ = 67.8 yrs, $M_{Older\_Slow}$ = 68.5 yrs; $t(30)$ = 0.37, $p$ = .72, $d$ = 0.13, BF = 0.36), indicating that learning differences within the Older group cannot simply be due to within group age variation.

#### *3.3.3.3 VALMT delayed cued-recall performance*

To investigate delayed recall performance and forgetting the mean recall scores are plotted below. Figure 3.8 shows the delayed recall performance of the Younger group and the combined Older group, while Figure 3.9 shows the performance for the Younger, Older_Fast, and Older_Slow groups.

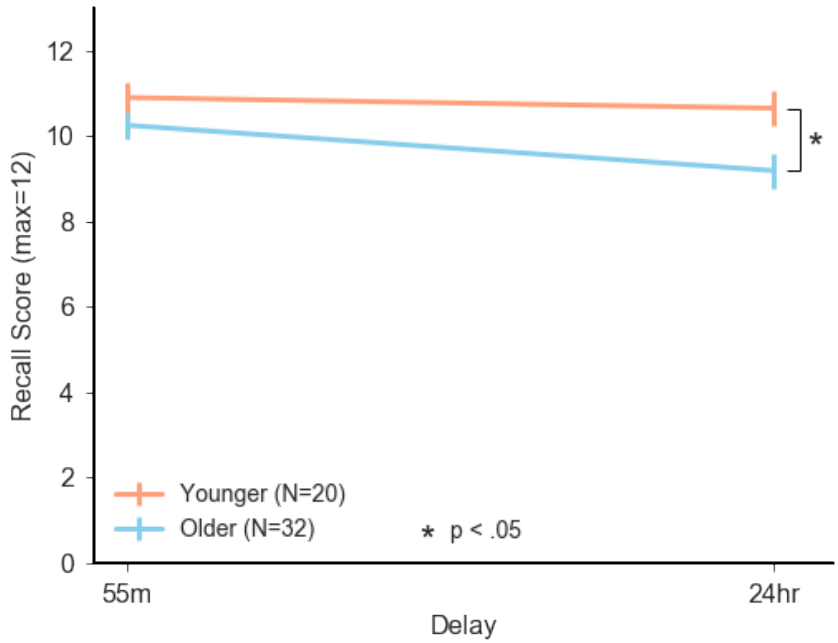


*Figure 3.8 Delayed recall performance at 55min and 24hr delays for the Younger and combined Older groups (error bars +/- 1SE).*

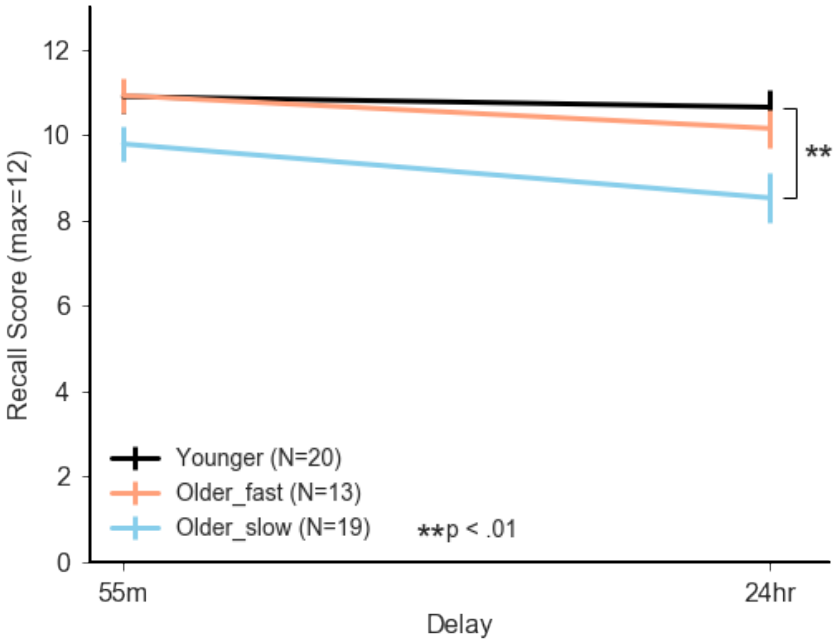


*Figure 3.9 Delayed recall performance at 55min and 24hr delays for the Younger and fast and slow learning Older groups (error bars +/- 1SE).*

#### 3.3.3.3.1 Combined Older group

The variation in cued recall performance between groups and across delay intervals was analysed using a mixed factors ANOVA with within-subjects factor Delay (55min vs 24hr) and between-subjects factor Age (Younger vs Older). There were significant main effects of Age ($F(1, 50) = 4.31$, $p = .043$, $\eta_p^2 = 0.07$, BF = 3.55) and Delay ($F(1, 50) = 10.43$, $p = .002$, $\eta_p^2 = 0.03$, BF = 61.4) and an interaction that approached significance ($F(1, 50) = 3.99$, $p = .051$, $\eta_p^2 = 0.01$, BF = 5.2). This indicates there is a significant amount of forgetting between 55min and 24hrs overall and that compared to the Younger group the Older group had a statistically lower level of recall overall (Marginal means: $M_{Older} = 9.72$ pairs, $M_{Younger} = 10.77$ pairs) and a forgetting rate between 55min and 24hr which may indicate higher forgetting in older people in the general population.

To compare cued recall performance across groups at each time-point two independent sample t-tests were used. The Younger group scored higher than the Older group at both delays. While at 55min the effect size was small and the difference was not significant ($M_{Younger} = 10.9$ pairs , $M_{Older} = 10.2$ pairs, $t(50) = 1.34$, $p = .19$, $d = 0.38$, BF = 0.59), at 24hrs the effect size was medium and the difference had become significant ($M_{Younger} = 10.7$ pairs, $M_{Older} = 9.2$ pairs, $t(50) = 2.43$, Bonferroni adjusted $p = .04$, $d = 0.70$, BF = 2.96).

The forgetting rate between the 55min and 24hr time points was calculated as *[55min score – 24hr score]/24hr score.* The Older group had a higher forgetting rate than the Younger group, with a difference that was on the cusp of significance ($M_{Younger}$ = 0.02, $M_{Older}$ = 0.11; $t(50) = 1.96$, $p = .05$, $d = 0.56$, BF = 1.34), suggesting that older people may display faster forgetting over this interval.

#### 3.3.3.3.2 Fast and slow learning Older groups

Equivalent analyses of cued recall performance were performed with the Older group split into Fast and Slow learning subgroups. A mixed factors ANOVA with within-subjects factor Delay (55min vs 24hr) and between-subjects factor Group (Younger vs Older_Fast vs Older_Slow) identified significant main effects of Delay ($F(1, 49) = 14.25$, $p < .001$, $\eta_p^2 = 0.04$, BF = 49.2) and Group ($F(2,49) = 4.82$, $p = .012$, $\eta_p^2 = 0.14$, BF = 6.02), but no significant interaction ($F(2,49) = 2.46$, $p = .096$, $\eta_p^2 = 0.01$, BF = 2.67). Bonferroni adjusted post hoc tests for the main effect of Group found that while there was no significant difference between the Younger and Older_Fast group ($p = 1.00$, d = 0.13, BF = 0.29), there was a difference between the Older_Fast and Older_Slow groups that had medium effect size and approached significance (p = .09, d = 0.74, BF = 5.41), and a large and significant difference between the Younger and Older_Slow ($p = .015$, d = 0.87, BF = 48.5). Taking the Bayes factors into account, there is moderate evidence that young and fast learning older people in the population perform the same, while there is moderate to strong evidence that slow learning older people perform lower than both fast learning older people and young people.

Recall scores at each individual delay were compared using one-way ANOVAs, and significant results investigated using Bonferroni adjusted post hoc tests for pairwise comparisons. There was no significant difference between the group means at 55min ($F(2,49) = 2.73$, $p = .075$, $\eta_p^2 = 0.10$, BF = 1.07). However, by 24hr the difference was significant ($F(2,49) = 5.65$, $p = .006$, $\eta_p^2 = 0.19$, BF = 8.21). Post-hoc tests at this longer delay showed that while there was no significant difference between the Younger and Older_Fast groups' performance ($p = 1.00$, d = 0.24, BF = 0.44), the Older_Slow group scored below the Older_Fast and Younger groups, with a difference that approached significance for the Older_Fast comparison ($p = .093$, d = 0.80, BF = 1.67), and was large and significant for the Younger comparison ($p = .006$, d = 1.04, BF = 10.0).

A one-way ANOVA comparing forgetting rates found an overall difference in the Younger, Older_Fast and Older_Slow group means that approached significance ($F(2, 49) = 2.69$, $p = .08$, $\eta_p^2 = 0.10$, BF = 1.07). While Bonferroni adjusted post-hoc tests confirmed a small and non-significant difference between the Younger and Older_Fast groups ($p = 1.00$, d = 0.30, BF = 0.49) and a larger but still non-significant difference between Older_fast and Older_Slow ($p = .68$, d = 0.44, BF = 0.57), the difference between Older_Slow and Younger groups was medium and approached significance ($p = .07$, d = 0.74, BF = 1.89). This indicates that the Older_Slow group are driving the higher forgetting in the Older group.

#### 3.3.3.3.3 Evidence of accelerated forgetting between 55min and 24hr

Although there is evidence of greater forgetting between 55m and 24hr for the Older group, and in particular the Older_Slow group, these participants also make more errors during learning and score lower at 55min recall. This means their forgetting appears to start early. To provide the strongest evidence of late-onset ALF we would need evidence of participants who perform normally for learning and 55min recall, but then go on to show accelerated forgetting after 55min.

One approach to analysing this to identify those participants who are normal with respect to their age group on learning performance and 55min recall, but who then are impaired on forgetting rate between 55m and 24hr. To operationalise this, subsets of the Younger and Older groups were prepared consisting of those whose total trials to complete learning was no more than one SD above or below the mean for the relevant age group, and whose 55min recall score was no more than 1SD above or below the group mean. This resulted in a subsets containing 16 for the Younger group and 25 for the Older group

Anyone in these groups who then went on to score at least 1.5SDs below the group mean for forgetting rate between 55min and 24hrs was deemed to display late onset ALF. This threshold should identify those in the bottom 6.68% in an approximately normal distribution. None of the Younger or Older groups met this criteria. The sample sizes are small for this type of probabilistic analysis, however the results suggest there is no evidence of late-onset ALF in either the Younger or the Older group.

#### 3.3.3.3.4 Impact of variation in test delay

The inclusion criteria for the main analyses of online VALMT data include acceptable time windows for completion of the 55m and 24hr delayed tests. The window sizes are +/- 10mins for the 55min delay and +/- 4hrs for the 24hr delay. Experiment 3 validated these for Young participants. To validate these criteria with a group reflecting the entire adult age range this analysis was repeated, investigating the relationship between test delay and cued-recall score. To provide the widest possible range of delays this analysis was performed for the Delay_variation group, which includes those who completed their delayed tests outside the acceptable windows. This may also provide some indication of the likely impact of extending the 24hr delay to a longer interval. The scatterplots in Figure 3.10 illustrate the relationships for this group.

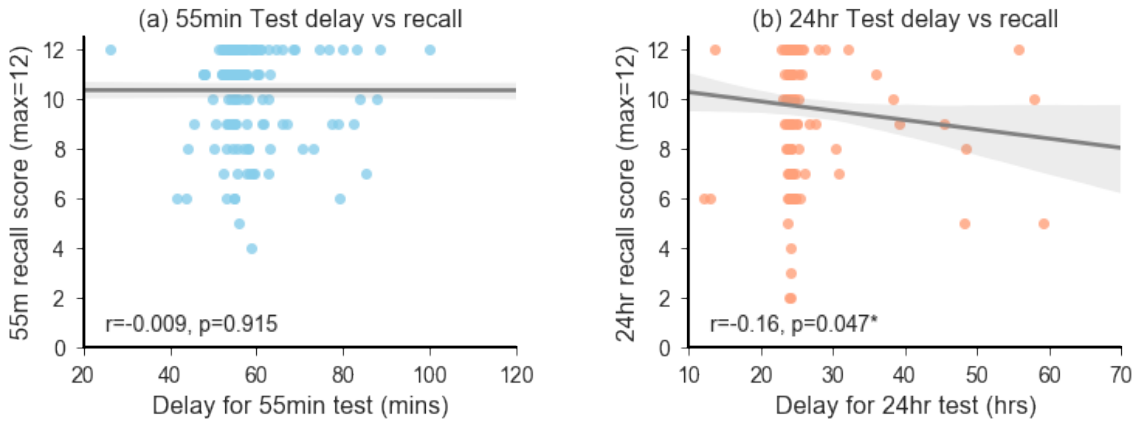


*Figure 3.10 Correlation of cued-recall scores with test delay, for (a) the 55m and (b) the 24hr delayed tests, for the Delay_variation group (N=154); shaded area is 95% confidence interval.*

There was negligible correlation between the observed test delay and recall score for the 55min test (r = .009, p = .91, BF = 0.10; see Figure 3.10a), replicating the result from Experiment 3. However, unlike Experiment 3 the correlation for the 24hr test, although small, was statistically significant (r = 0.16, p = .047, BF = 0.71; see Figure 3.10b). Visual inspection suggests this may be driven by the small number of participants who completed this test beyond 40 hours, so this result may not be relevant to validating the 24hr test window. To provide a more representative test the analysis was repeated using the All_Criteria_Met group, to investigate the relationship between delay and recall within the current window size. This is shown in Figure 3.11.

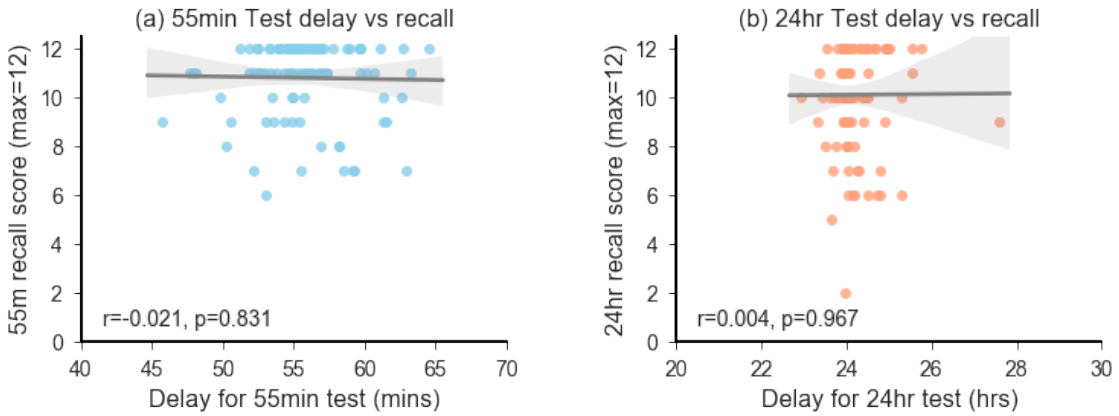


*Figure 3.11 Correlation of cued-recall scores with test delay, for (a) the 55m and (b) the 24hr delayed tests, for the All_Criteria_Met group (N=104); shaded area is 95% confidence interval.*

There was negligible correlation between the observed test delay and recall score for both the 55min test (r = .02, p = .583 BF = 0.125; see Figure 3.11a) and the 24hr test ((r = .004, p = .97, BF = 0.123; see Figure 3.11b). This suggests variation in the timing of the delayed tests within the acceptable windows is unlikely to have impacted recall scores, and the windows are therefore acceptable.

There were only three participants who completed their 24hr test at beyond 2 days so, as with Experiment 3, there is no clear evidence for the likely impact of extending the second test delay to three days or one week.

#### *3.3.3.4 Subjective memory complaints*

To illustrate how total MCS scores vary across groups Figure 3.12 compares the MCS scores of the All_Criteria_Met group, Younger group, combined Older group, and the Older group split into Fast_Older and Slow_Older subgroups.

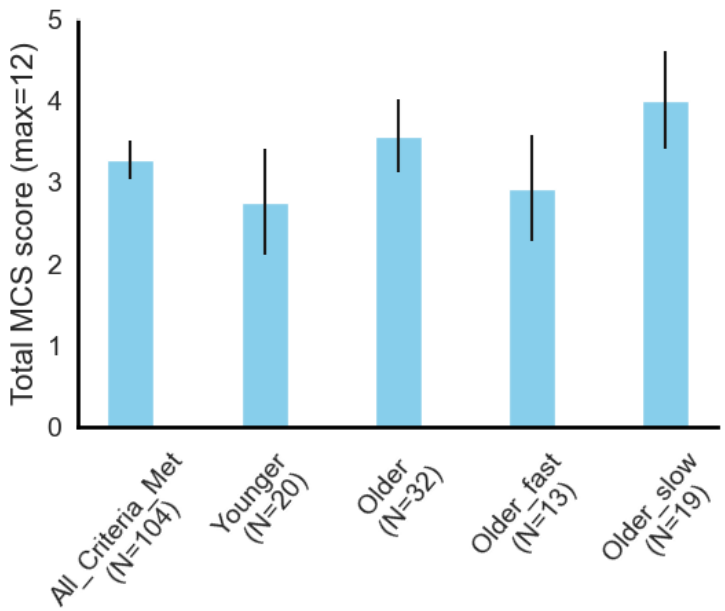


*Figure 3.12 Mean total MCS score as a function of group (error bars +/- 1SE).*

#### 3.3.3.4.1 Combined Older group

To investigate group differences in memory complaints the MCS scores for Younger and Older groups were compared using an independent sample t-test. The Older group reported more complaints but this difference was not statistically significant ($M_{Older}$ = 3.56, $M_{Younger}$ = 2.75, $t$ (50) = 1.06, $p$ = .29, $d$ = 0.30, BF = 0.45). This indicates the total MCS score alone cannot account for the group differences seen in learning and memory performance.

#### 3.3.3.4.2 Fast and slow learning Older groups and MCS severity

To investigate the impact on group MCS differences of categorizing the Older participants as fast or slow learners the MCS scores for the Younger, Older_Fast and Older_Slow groups were compared using a one-way ANOVA. There was no significant difference between the three group scores ($F$(2,49) = 1.19, $p$ = .31, $\eta_p^2$ = 0.05, BF = 0.36), indicating no significant effect of Group on memory complaints as measured by the total MCS score.

To investigate the impact of MCS severity the distribution of memory complaints within each group is summarized in Table 3.5, where scores are categorized into four ordinal categories using an adjusted version of the scoring scheme used in Experiment 2: No Memory Complaints (MCs: 0-2), Mild MCs (3-5), Moderate MCs (6-8), Severe MCs (9-12). For comparison the breakdown is also reported for older participants who failed to

learn all pairs to criterion within the time limit (Older_Exc, N=7) but met all other inclusion criteria.

*Table 3.5 Distribution of Memory Complaints as a function of group.*

| Memory Complaints Category | Younger N=20 | Older N=32 | Older_Fast N=13 | Older_Slow N=19 | Older_Exc N=7 |
|---|---|---|---|---|---|
| None | 14 (70%) | 14 (44%) | 8 (62%) | 6 (32%) | 1 (14%) |
| Mild | 2 (10%) | 7 (22%) | 2 (15%) | 5 (26%) | 3 (43%) |
| Moderate | 3 (15%) | 10 (31%) | 3 (23%) | 7 (37%) | 3 (43%) |
| Severe | 1 (5%) | 1 (3%) | 0 (0%) | 1 (5%) | 0 (0%) |

While 70% of the Younger and 62% of Older_Fast groups reported no memory complaints, only 32% of the Older_Slow did. This indicates that while there are no significant group differences in total MCS score, group differences are seen once MCS severity is explicitly taken into account. Only 14% of the Older_Exc group reported no problems, suggesting those who struggle to learn to criterion also suffer from more memory problems.

#### 3.3.3.4.3 Relationship between subjective memory complaints and VALMT delayed recall

To investigate how subjective memory complaints related to objective recall performance on VALMT these variables are plotted as scatterplots in Figure 3.13, for the All_Criteria_Met group which includes participants of all ages.

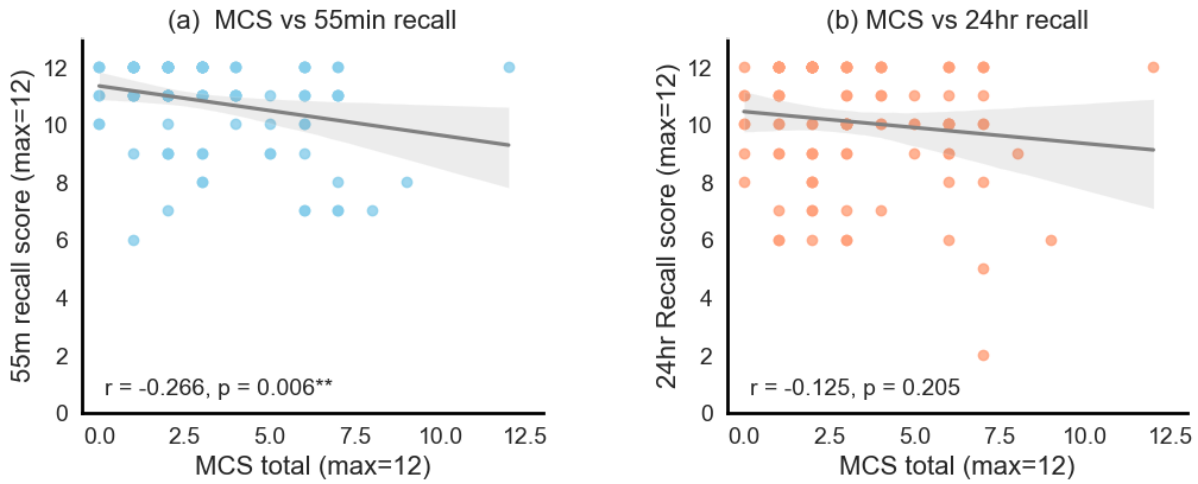


*Figure 3.13 Correlation of memory complaints with delayed recall scores at (a) 55m and (b) 24hr for the All_Criteria_Met group (N=104) ); shaded area is 95% confidence interval.*

While there was a small but significant correlation between memory complaints and VALMT recall score at 55mins (r = -0.266, p = .006, Bonferroni adjusted p = .012, BF =

4.76; see Figure 3.13a), there was no significant correlation at 24hr (r = -.125, p=.205, BF = 0.27; see Figure 3.13b).

#### 3.3.3.4.4 Relationship between age and learning errors

To investigate how age related to the number of errors made during learning these variables are plotted as scatterplots in Figure 3.14. The relationship is first shown for the All_Criteria_Met group and then for the Older group, to investigate whether the greater variation in learning performance seen in older participants may be due to variation in age within that group.

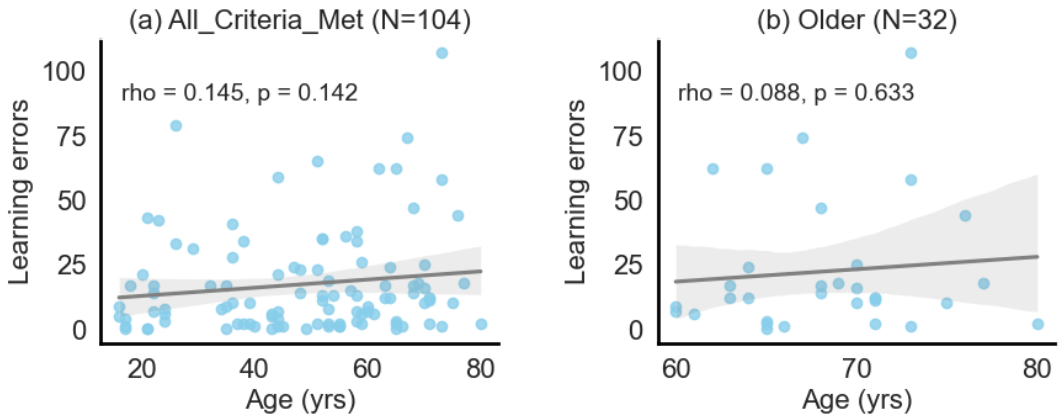


*Figure 3.14 Correlation of age with learning errors for (a) the All_Criteria_Met group and (b) the Older group alone; shaded area is 95% confidence interval.*

There was no significant correlation between age and learning errors for the All_Criteria_Met group (rho = .14, p = .16; see Figure 3.14a) or the Older group (rho = .09, p=.63; see Figure 3.14b), indicating age is not a major driver of learning performance, especially in the Older group.

### *3.3.3.5 Relationship between speed of response and learning errors*

Making more errors during learning may be due to being less familiar with using computers, especially in the Older group. One proxy for this familiarity is speed of response, as it is expected that someone who is less familiar with computers will take longer over each attempt. To investigate the relationship these variables are plotted as scatterplots in Figure 3.15, for the All_Criteria_Met group and separately for the Older group.

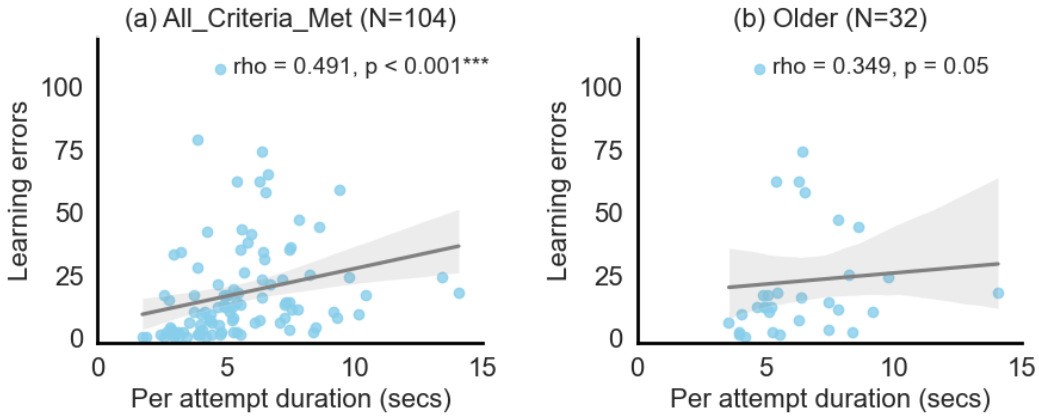


*Figure 3.15 Correlation of learning errors with speed of response (duration per attempt) for (a) the All_Criteria_Met group and (b) for the Older group alone; shaded area is 95% confidence interval.*

There was a positive correlation between learning errors and per-attempt duration for both groups, with those that take longer over each attempt making more errors. While this correlation is strongly significant for the All_Criteria_Met group (rho = 0.49, p < .001; see Figure 3.15a), it was weaker for the Older group and was on the cusp of significance (rho = 0.35, p = .05; see Figure 3.15b).

The slope of the line of best fit in Figure 3.15b (the change in errors associated with a change in duration) suggests that the even if per-attempt duration is related to the number of errors made, its impact is small. In addition, a direct comparison shows that the Older_Slow group do not take significantly longer per-attempt than the Older_Fast group ($M_{Older\_Fast}$ = 5.75secs, $M_{Older\_Slow}$ = 6.68secs, t(30) = 1.19, p = .24, d = 0.43, BF = 0.58). Together, this suggests this factor is unlikely to be a major driver of the greater variation in learning performance seen in older participants ($Mdn_{Fast\_Older}$ = 3.0 errors, $Mdn_{Slow\_Older}$ = 18.0 errors).

### *3.3.3.6 Relationship between memory complaints and learning errors*

To investigate how memory complaints related to learning performance the number of errors made during learning is plotted against MCS score in Figure 3.16, for the All_Criteria_Met group, and separately for the Older group to investigate whether the greater variation in learning performance seen in older participants may be associated with variation in memory complaints.

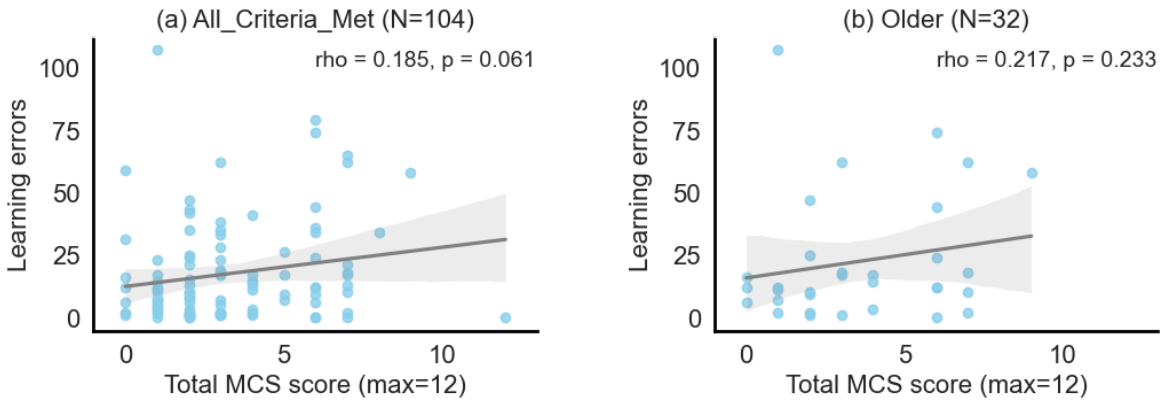


*Figure 3.16 Correlation of learning errors with memory complaints for (a) the All_Criteria_Met and (b) the Older group alone; shaded area is 95% confidence interval.*

For both groups there was a small positive correlation between learning errors and memory complaints, with those who report more complaints also making more errors. While this approached significance for the All_Criteria_Met group (rho = 0.19, p = .06; see Figure 3.16a) it was not significant for the Older group (rho = 0.22, p = .31; see Figure 3.16b) although this difference in p-value may be due to the smaller sample size. This suggests this factor is unlikely to be a major driver of the greater variation in learning performance seen in older participants

#### *3.3.3.7 Distribution of learning errors across word-pairs and relationship to recall*

To investigate how delayed recall and errors made during learning vary across the 12 word-pairs, and to illustrate how these two variables relate to each other, Figure 3.17 shows the mean number of errors made for each individual word-pair plotted against the corresponding mean recall score at each delay, for the All_Criteria_Met group and separately for the Older group.

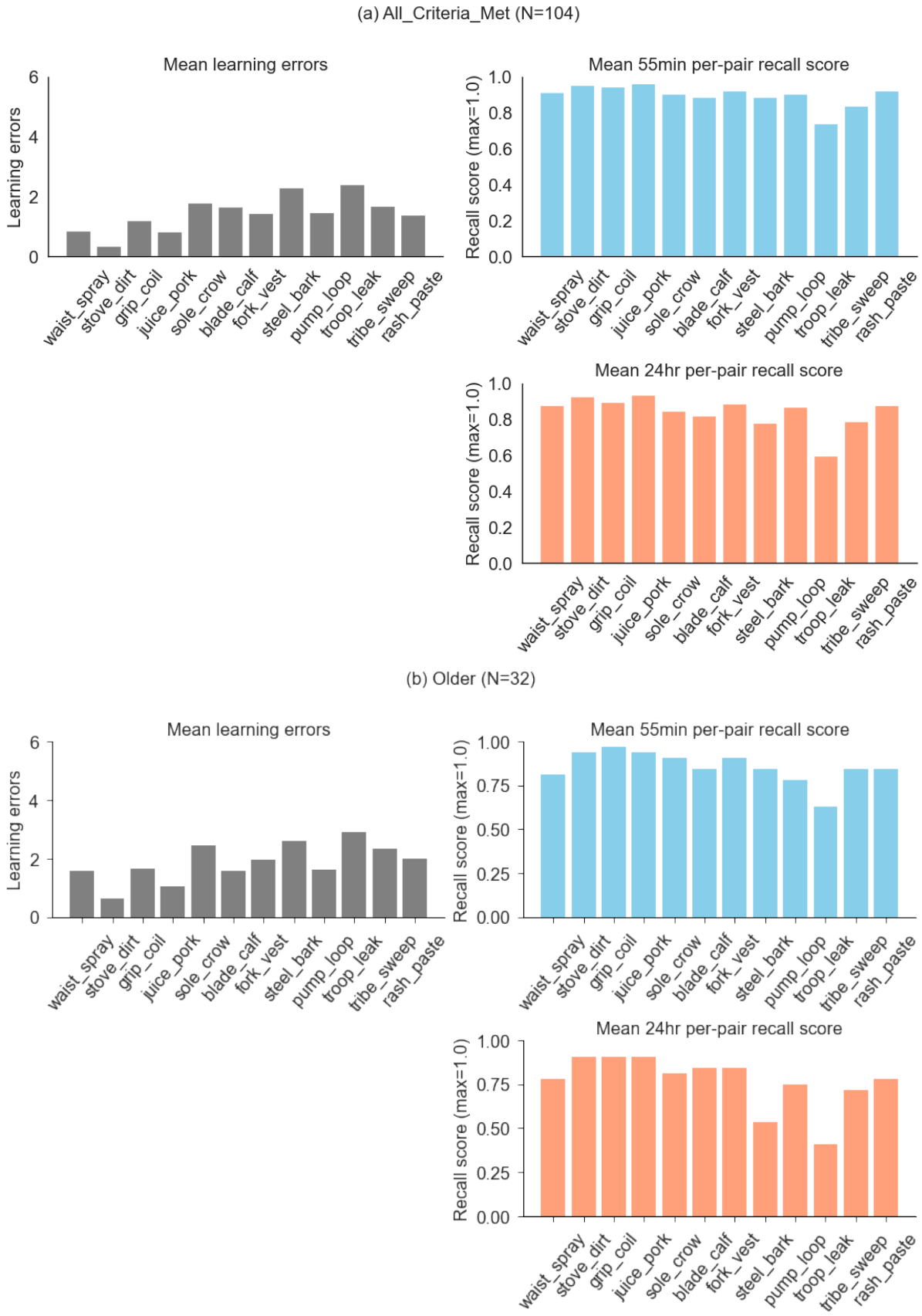


*Figure 3.17 Mean learning errors and delayed recall by individual word-pair for (a) the All_Criteria_Met group and (b) the Older group.*

Visual inspection shows that the distribution of errors and recall across pairs is very similar for both groups. Those pairs which generate the fewest errors are recalled best while those pairs which resulted in most errors are recalled most poorly. There is variation in difficulty between pairs. This will be beneficial for distinguishing between performance levels, since some more difficult pairs are needed to challenge and differentiate between higher performers, and easier ones are needed to help differentiate between lower performers. No pairs encounter zero or a very high number of errors, or zero or 100% recall, so all pairs are providing useful information.

#### *3.3.3.8 Relationship between delayed recall and total and per-pair learning errors*

To investigate how the total number of errors made during learning related to subsequent delayed recall the mean number of learning errors is plotted against recall score in Figure 3.18, for the All-Participants group and separately for the Older group.

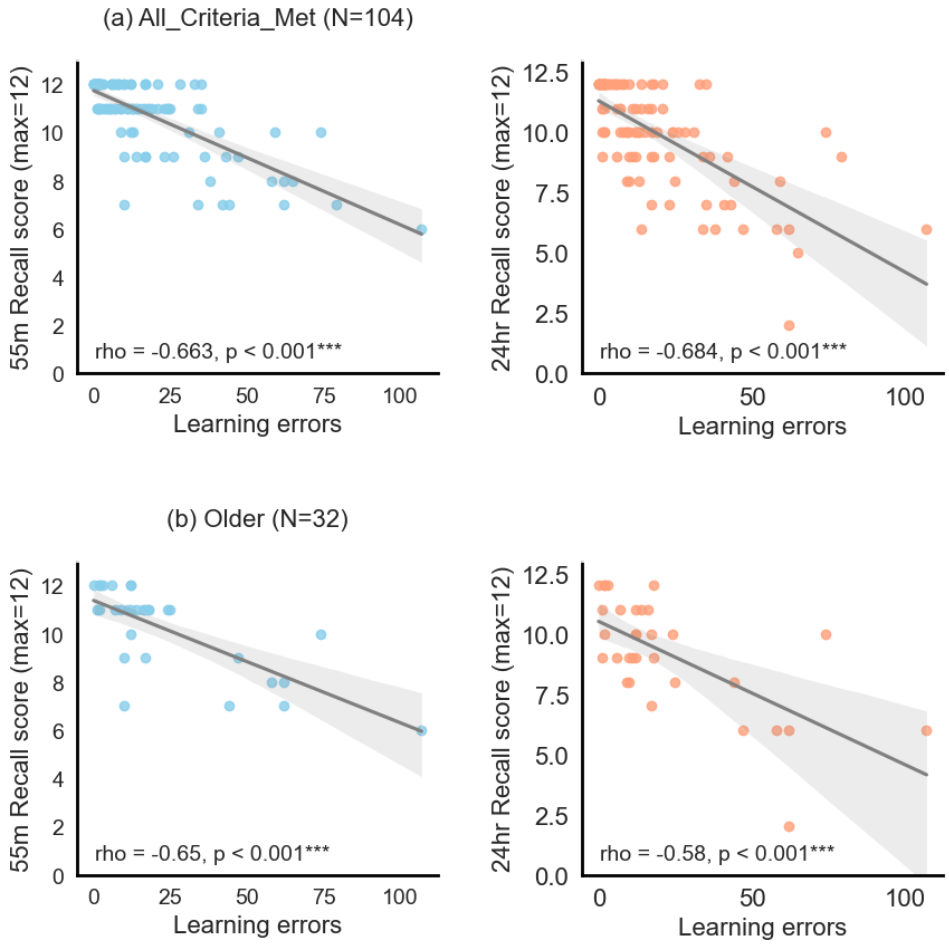


*Figure 3.18 Correlation of learning errors with delayed recall at 55m (left) and 24hr (right) for (a) All_Criteria_Met and (b) the Older group alone; shaded area is 95% confidence interval.*

There was a large and significant negative correlation between learning errors and delayed recall at both 55min and 24hr for both the All_Criteria_Met group (55min rho = -0.66, p < .001; 24hr rho = -0.68, p < .001; see Figure 3.18a) and the Older group (55min rho = -0.65, p < .001; 24hr rho = -0.58, p < .001; see Figure 3.18b). This indicates a strong link between learning errors and subsequent recall, with those participants who make more errors scoring lower at recall.

To investigate this association at the more granular per-pair level, the scatterplots in Figure 3.19 illustrate the relationship between the mean number of errors made learning each individual pair and the mean recall scores for that specific pair. These plots show one data-point for each of the 12 word-pairs.

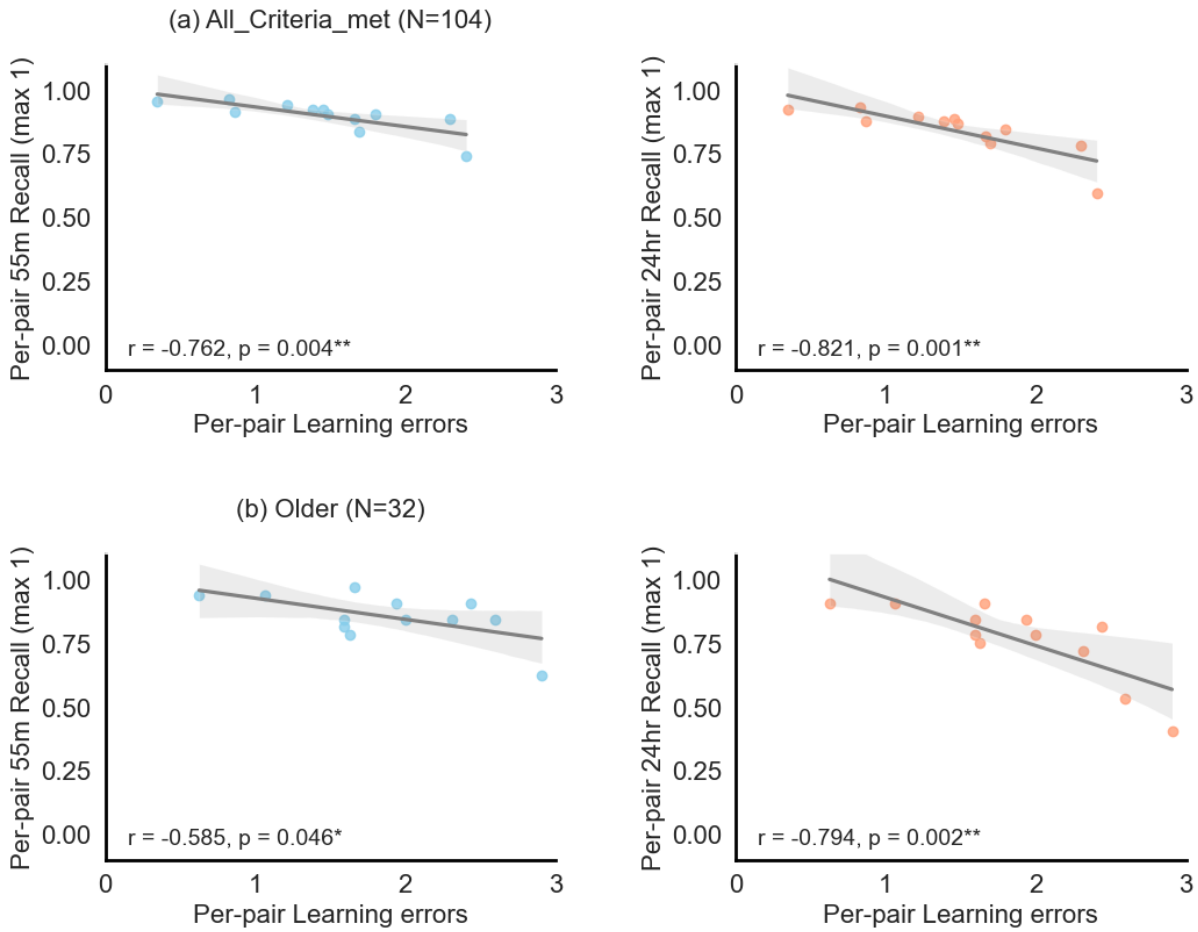


*Figure 3.19 Correlation of learning errors per word-pair with delayed recall per word-pair at 55m (left) and 24hr (right) for (a) All_Criteria_Met (upper) and (b) the Older group alone (lower); shaded area is 95% confidence interval.*

There was a statistically significant and large or very large negative correlation between learning errors per individual word-pair and delayed recall per word-pair at both 55min and 24hr for the All_Criteria_Met group (55min r = -0.762, p = .004, BF = 14.3;

24hr r = -0.821, p = .001, BF = 40.7; see Figure 3.19a) and for the Older group (55min r = -0.585, p = .046, BF = 2.13; 24hr r = -0.794, p = .002, BF = 24.4; see Figure 3.19b). Overall, this analysis confirms the visual pattern seen in Figure 3.17, in which those pairs which encounter most errors are recalled most poorly.

#### *3.3.3.9 Recall after errorless versus errorful learning*

Participants typically complete learning to criterion for some pairs without making any errors, while for other pairs they do make some errors. To further investigate the relationship between learning errors and subsequent forgetting each participant's recall scores for pairs learnt with no errors (*errorless*) and with errors (*errorful*) were separated, and the group means for these were analysed.

To investigate whether errorful pairs are forgotten faster, recall for both errorful and errorless pairs are illustrated in Figure 3.20 for the All_Criteria_Met group and the Older group. To help interpret these forgetting curves the mean number learnt to criterion in each way (errorless, errorful) is shown.

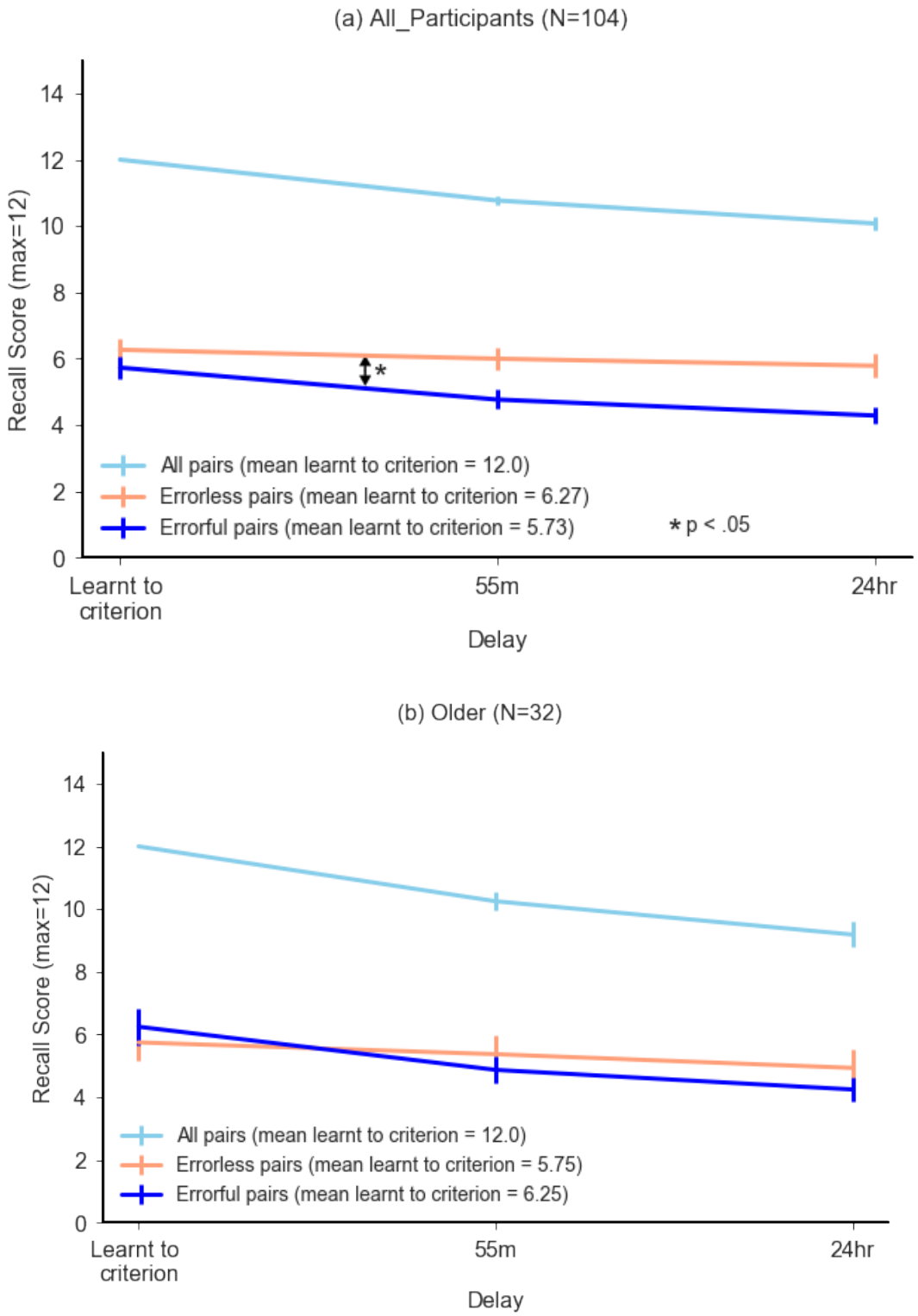


*Figure 3.20 Delayed recall performance for word-pairs learnt with and without errors (errorful, errorless) for (a) the All_Criteria_Met and (b) the Older groups.*

For both groups the mean number of pairs learnt errorlessly and errorfully are similar, with approximately half of the total pairs learnt in each manner (All_Criteria_Met group:

errorless mean 6.27, errorful mean 5.73; Older group: errorless mean 5.75, errorful mean 6.25).

To investigate forgetting two rates were analysed: *early forgetting*, defined as forgetting between completion of learning to criterion and 55min; *late forgetting*, defined as forgetting between 55min and 24hr. In each case a relative rate was used, using the standard formula: [score at first delay – score at second delay] / score at first delay. These rates were compared for errorless and errorful pairs. Dependent means t-tests for the All_Criteria_Met group found the forgetting of errorful pairs was greater than that for errorless pairs, for both early and late forgetting. While the difference was very small and non-significant for late forgetting ($M_{errorless}$ = 0.047, $M_{errorful}$ = 0.053, $t$ (92) = .03, $p$ = .97, $d$ = 0.004, BF = 0.11), it was medium and significant for early forgetting ($M_{errorless}$ = 0.06, $M_{errorful}$ = 0.13, $t$ (93) = 3.05, $p$ = .003, $d$ = 0.32, BF = 8.42). Equivalent analysis for the Older group found a very similar pattern; the difference was very small and non-significant for late forgetting ($M_{errorless}$ = 0.11, $M_{errorful}$ = 0.09, $t$ (28) = .84, $p$ = .41, $d$ = 0.16, BF = 0.27), and was larger for early forgetting, although this only approached significance despite the same effect size as the All_Criteria_Met group, probably due the smaller sample size ($M_{errorless}$ = 0.09, $M_{errorful}$ = 0.17, $t$ (29) = 1.72, $p$ = .09, $d$ = 0.32, BF = 0.72). Taken together, these results suggests that errorful pairs suffer greater forgetting than errorless pairs to 55mins, but similar forgetting after that.

Previous results have shown that slow learning older participants make more errors, and also score lower at delayed recall. Their lower recall performance could be due to a greater underlying forgetting rate, or could be due to interference from the errors, or a combination of these factors. One way to investigate this is to analyse forgetting for errorless pairs, where any interference from errors should be greatly reduced. To illustrate group differences in underlying forgetting rates the delayed recall scores for errorless pairs for Younger, Older_Fast and Older_Slow groups are illustrated in Figure 3.21.

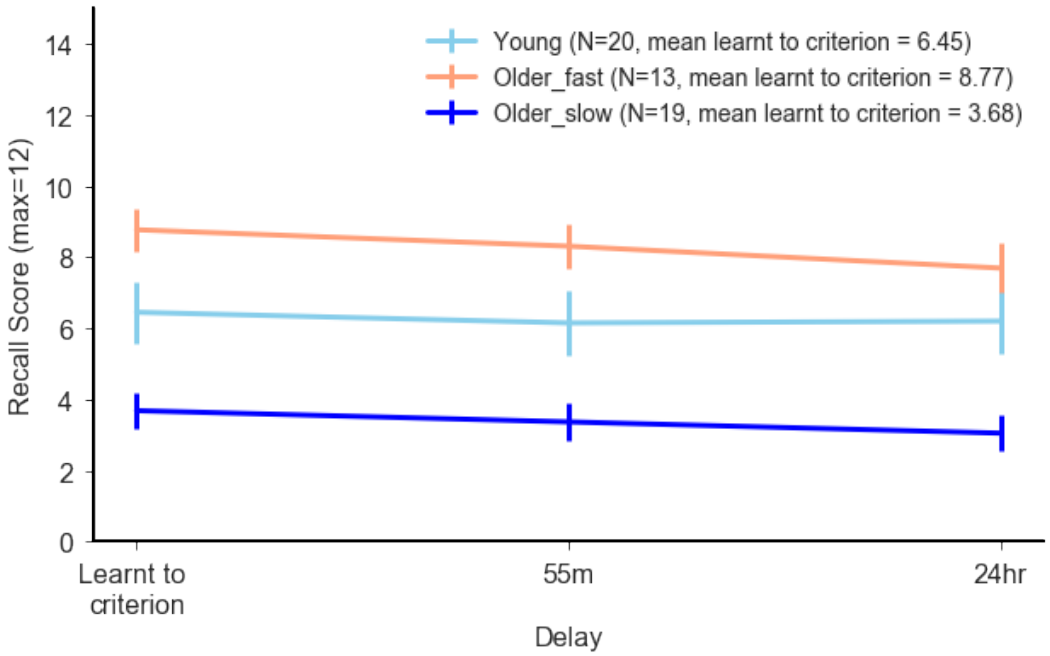


*Figure 3.21 Delayed recall performance for word-pairs learnt without errors (errorless) for the Younger, Older_Fast and Older_Slow groups (error bars +/- 1SE)*

A one-way ANOVA comparing the number of pairs learnt errorlessly by each group found a significant difference between means ($F(2, 49) = 11.84$, $p < .001$, $\eta_p^2 = 0.33$, BF = 345). Bonferroni adjusted post hoc tests found that while there was no significant difference between the Younger and Older_Fast groups (p = .096, d = 0.79, BF = 1.46), the Older_Slow group learnt significantly fewer pairs errorlessly than both the Younger (p = .015, d = 0.94, BF = 5.08) and Older_Fast groups ($p < .001$, d = 1.73, BF = 23636).

One-way ANOVAs comparing the early and late forgetting rates across groups found no significant difference between early forgetting rates ($F(2, 47) = .48$, $p = .62$, $\eta_p^2 = 0.02$, BF = 0.34) but a difference in late forgetting rates that approached significance ($F(2, 46) = 3.11$, $p = .054$, $\eta_p^2 = 0.12$, BF = 0.50). Bonferroni adjusted post hoc tests for late forgetting rates found no significant differences between the Older_Slow group and Older_Fast groups (p = 1.00, d = 0.35, BF = 0.95), and while both Older groups forgot more than the Younger group this difference was not significant for the Older_Fast group (p = .57, d = 0.48, BF = 0.35) but approached significance for the Older_Slow group (p = 0.050, d = 0.83, BF = 0.74).

In this type of experiment the standard relative forgetting formula is intended to adjust for differences in performance at the first interval, by assessing forgetting as a proportion of the number recalled at the first interval; in this case that means allowing comparisons to be made between groups who learnt different numbers of pairs to criterion. This is an

accepted approach where the differences are small, but in this case the difference in the number of pairs learnt to criterion is large for the Older_Fast and Older_Slow groups (8.77 vs 3.68). An alternative approach sometimes used is to use the absolute difference in scores, without any adjustment for the performance at the initial time point. As a precaution, the ANOVAs comparing early and late forgetting rates were repeated, but using the absolute difference formula: [score at first delay – score at second delay]. This again found no significant difference between early forgetting rates ($F(2, 47) = .20$, $p = .82$, $\eta_p^2 = 0.01$, BF = 0.18) but a difference in late forgetting rates which in this case did reach significance ($F(2, 47) = 3.84$, $p = .03$, $\eta_p^2 = 0.14$, BF = 2.21). Bonferroni adjusted post hoc tests for late forgetting rates again found no significant differences between the Older_Slow group and Older_Fast groups (p = .79, d = 0.41, BF = 0.49), and while both Older groups forgot more than the Younger group this difference was not significant for the Older_Slow group (p = .28, d = 0.56, BF = 1.67) but was significant for the Older_Fast group (p = 0.03, d = 0.98, BF = 4.74).

Taken together, these relative and absolute formulae suggest Older participants have a similar underlying forgetting rate to Younger participants to 55min, and may have a slightly faster forgetting rate between 55min and 24hr. However, importantly, neither relative nor absolute forgetting rate analysis show any evidence that Older_Slow participants have a faster underlying forgetting rate than Older_Fast participants over either time period.

#### *3.3.3.10 Characteristics of participants who failed to learn to criterion*

There were 15 participants who failed to learn all 12 word-pairs to criterion within the fixed 20min time limit (*expired*) but met all the other inclusion criteria. In the original face-to-face version of the VALMT these participants would have been allowed to continue learning for as long as they required. To help understand the impact of excluding these expired participants their characteristics were analysed.

Of the 15 only one was aged under 30, while seven were aged over 60yrs, so the exclusions will have had minimal impact on the Younger group but could have impacted the Older group. Of the seven aged 60yrs+, six reported memory complaints; the mean MCS score for the seven was 4.86, compared to 3.56 for the Older group who completed learning to criterion. So, excluding them will have reduced the mean MCS score for the Older group. Although they did not complete learning to criterion within the 20min limit ,

by the end of the 20min period their mean learning errors total was already 71.8, compared to 22.3 for the Older group who completed learning to criterion successfully. Excluding them will therefore have reduced the mean trials to criterion for the Older group.

Including these previously excluded participants back into the analysis for learning performance and memory complaints increased the group differences in the Younger vs Older comparisons for these variables. The relevant statistical test results with and without these expired participants are compared below. Note that it was not possible to include these expired participants back into analyses of delayed recall as, by definition, they did not manage to learn all pairs to criterion.

Comparison of trials to complete learning:
Expired excluded: $Mdn_{\text{Older}} = 48.0$, $Mdn_{\text{Younger}} = 44.5$;
$MWU = 370$, $p = .35$, $r = 0.16$, BF = 0.44
Levene's test of equality of variances not violated:
Older $\sigma^2 = 648.8$, Younger $\sigma^2 = 401.0$, p = .35
Expired included: $Mdn_{\text{Older}} = 53.0$, $Mdn_{\text{Younger}} = 44.5$;
$MWU = 498$, $p = .085$, $r = 0.28$, BF = 1.00
Levene's test of equality of variances is violated:
Older $\sigma^2 = 864.1$, Younger $\sigma^2 = 401.0$, p = .02

Comparison of memory complaints:
Expired excluded: $M_{Older} = 3.56$, $M_{Younger} = 2.75$,
$t(50) = 1.06$, $p = .29$, $d = 0.30$, $BF = 0.45$
Expired included: $M_{Older} = 3.79$, $M_{Younger} = 2.75$,
$t(57) = 1.44$, $p = .15$, $d = 0.40$, $BF = 0.65$

The correlation between memory complaints and learning errors within the Older group also increased when these previously excluded participants were included in the analysis.

Comparison of correlation between memory complaints and learning errors:
Expired excluded: rho = 0.217, p = .23
Expired included: rho = 0.281, p = .03

Including the expired participants back into these analyses had a significant effect, increasing group differences enough to change the result of two statistical tests. First, the variance in the Older groups learning performance becomes statistically greater than the variance of the Younger group. Second, the correlation between memory complaints and learning performance for the Older group becomes significant. Their inclusion also moved the Older vs Younger group comparison for learning trials close to significance (p

changed from .35 to .085), and if these expired participants had been allowed to continue learning (as they were in the face-to-face version) they would have made further errors which may have resulted in this comparison becoming significant too.

An alternative way to analyse this issue is to repeat the analysis for the data from Experiment 2, but exclude any participant who took longer than 20mins to complete learning during the first learning session; this simulates the effect of have an enforced 20min time limit. There were four participants who met this exclusion criteria, all in the Older group. A comparison of the relevant statistical tests is shown below:

Expt 2 Comparison of 55m cued recall:

Before exclusions: 55mins: $M_{\text{Older}}$ = 6.13 pairs, $M_{\text{Younger}}$ = 10.03 pairs, $t$ (43.0) = 5.03, Bonferroni adjusted $p$ <.001, $d$ = 1.32, BF = 3137

After exclusions: 55mins: $M_{\text{Older}}$ = 6.85 pairs, $M_{\text{Younger}}$ = 10.03 pairs, $t$(37.3) = 4.09, Bonferroni adjusted $p$ <.001, $d$ = 1.12, BF = 264

Expt 2 Comparison of trials to complete learning:

Before exclusions: 55mins: $Mdn_{\text{Older}}$ = 56.50 trials, $Mdn_{\text{Younger}}$ = 48.50 trials; $MWU$ = 194.5, $p$ < .001, $r$ = 0.57, BF = 52.26
Levene's test of equality of variances violated:
Older $\sigma^2$ = 1253.6, Younger $\sigma^2$ = 107.2, p = .005

After exclusions: 55mins: $Mdn_{\text{Older}}$ = 55.50 trials, $Mdn_{\text{Younger}}$ = 48.50 trials; $MWU$ = 585.5, $p$ = .001, $r$ = 0.50, BF = 7.46
Levene's test of equality of variances not violated:
Older $\sigma^2$ = 263.7, Younger $\sigma^2$ = 107.2, p = .22

Expt 2 Comparison of memory complaints:

Before exclusions: *$M_{Older}$ = 4.37, $M_{Younger}$ = 1.93, $t$ (48.6) = 3.73, $p$ = .001, $d$ = 0.98, BF = 61.79*

After exclusions: *$M_{Older}$ = 3.88, $M_{Younger}$ = 1.93, $t$ (41.5) = 2.90, $p$ = .006, $d$ = 0.79, BF = 9.51*

Comparison of correlation between memory complaints and learning errors:

Before exclusions: rho = 0.453, p < .001

After exclusions: rho = 0.310, p = .02

Excluding these participants reduced the Younger vs Older group differences as expected, due to the Older group's recall and learning performance both improving while their reported memory complaints reduced. For trials required to complete learning, the variance in the Older group was no longer statistically larger than that of the Younger

group (p changed from .005 to .22). For the remaining comparisons the change was not enough to flip the significance of any statistical test, and the effect sizes remained large.

Overall, these analyses suggest that while exclusions due to the 20min time limit do impact results, making Younger vs Older group differences smaller, this factor alone cannot account for all of the difference seen in the Young vs Older comparisons between the two versions of VALMT, so some other causal factor, or factors, must be present.

#### *3.3.3.11 Cross experiment comparisons*

To compare the online and face-to-face VALMT versions, Table 3.6 shows the number of trials needed to learn to criterion, the 55m delayed recall scores, the MCS total and the percentage reporting no complaints in this experiment and the equivalent numbers from Experiment 2, for both Younger and Older groups (it was not possible to compare 24hr results as that was not a delay in Experiment2). The number of trials taken during the first learning period of Experiment 2 is reported, rather than the mean across all three learning periods, to provide the best comparison with the online VALMT which has only one learning period.

*Table 3.6 Comparison of key face-to-face and online VALMT variables*

| Factor | Expt 2 Face-to-face *Mean(SD)* | Expt 4 Online *Mean(SD)* | Statistical comparison |
|---|---|---|---|
| **Younger** | | | |
| Trials to criterion (Younger) | Median 51.0, IQR 16.75 | Median 44.5, IQR 19.75 | $MWU = 330$, $p = .56$, $r = 0.10$, BF = 0.35 |
| 55m recall (Younger) | 10.03(1.89) | 10.90(1.61) | $t(48) = 1.65$, $p = .11$, $d = 0.48$, BF = 0.86 |
| No memory complaints | 80% | 70% | |
| MCS total | 1.83(1.64)* | 2.75(2.90) | $t(48) = 1.43$, $p = .16$, $d = 0.41$, BF = 0.65 |
| **Older** | | | |
| Trials to criterion (Older) | Median 68.0, IQR 33.25 | Median 48.0, IQR 17.5 | $MWU = 779$, $p < .001$, $r = 0.62$, BF = 90.98 |
| 55m recall (Older) | 6.13(3.72) | 10.25(1.71) | $t(60) = 5.56$, $p < .001$, $d = 1.41$, BF = 19551 |
| No memory complaints | 40% | 44% | |
| MCS total | 3.80(2.45)* | 3.56(2.54) | $t(60) = 0.37$, $p = .70$, $d = 0.09$, BF = 0.27 |

*MCS total for Expt 2 has been adjusted to match the coding scheme used in the online VALMT

The online and face-to-face Younger groups were matched on Gender ($X^2(1) = 0.62$, $p = .43$, BF = 0.49) and Language ($X^2(1) = 1.75$, $p = .19$, BF = 0.79), but were significantly different on Age, with the face-to-face Younger group being on average 3.5years older ($Age_{Online} = 21.3$, $Age_{Face\text{-}to\text{-}face} = 24.83$, $t(48) = 3.82$, $p < .001$, $d = 1.10$, BF = 69.6). Although statistically significant, a difference of 3.5yrs at this age should not cause a significant difference in cognitive function. The education coding scheme for the online VALMT is more granular than the face-to-face scheme, with eight categories instead of four. However, when the eight categories are collapsed into the four category scheme the groups appear matched ($X^2(3) = 4.41$, $p = .22$, BF = 0.75). The groups were also matched on total MCS scores, and the percentage reporting no memory complaints was also similar across versions. Overall, it appears that the groups are well-matched and cross-study comparisons for VALMT metrics will be valid.

Equivalent comparisons for the online and face-to-face Older groups showed the groups were matched on Gender ($X^2(1) = 0.007$, $p = .93$, BF = 0.33), Language ($X^2(1) = 2.95$, $p = .09$, BF = 1.58) and Education ($X^2(3) = 0.89$, $p = .83$, BF = 0.10), but were significantly different on Age, with the online Older group being on average 5.3 years older ($Age_{Online} = 68.25$, $Age_{Face\text{-}to\text{-}face} = 63.96$, $t(60) = 4.16$, $p < .001$, $d = 1.06$, BF = 219). The groups were also matched on total MCS scores, and the percentage reporting no memory complaints was also similar across versions. Overall, with the exception of an age difference, it appears that the groups are well-matched and cross-study comparisons for VALMT metrics will be valid. It should be noted that it is the online group who are older so the age difference, even if statistically significant, cannot account for the better performance of the online group on some VALMT metrics (as older participants would be expected to perform more poorly).

For the Younger groups there was no significant difference between the trials to criterion or 55m recall, suggesting the two versions operate in a similar manner for young participants.

However, for the Older group there was a large and significant difference for both variables with the online participants performing better, requiring fewer trials and recalling more pairs. This is a significant difference which requires interpretation (refer to the Discussion, Section 3.3.4.1).

### *3.3.3.12 Memory performance across the lifespan*

This is the first VALMT study to include participants across the entire adult age range. To investigate how delayed recall related to age these variables are plotted against each other in Figure 3.22, for the All_Criteria_Met group.

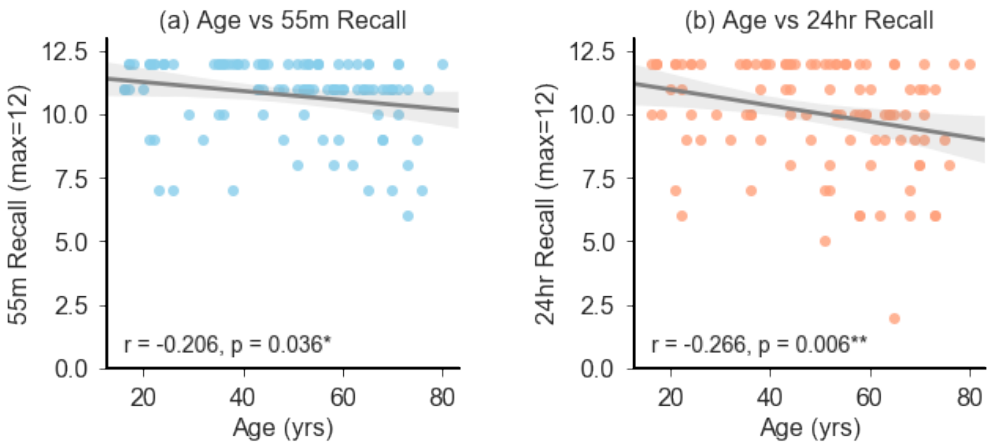


*Figure 3.22 Correlation of age with delayed recall at 55min and 24hr for the All_Criteria_Met group (N = 104); shaded area is 95% confidence interval.*

There was a small negative correlation between age and delayed recall which was significant at both 55min (r = -.21, p = .036, BF = 1.08) and 24hr (r = -.27, p = .006, BF = 4.78), indicating that older people score lower on delayed recall at both delays.

However, correlation analysis quantifies linear relationships between variables, and will miss any non-linear relationship. For example, perhaps memory performance is largely unchanged up to a certain age and then deteriorates at an increasing rate; correlation is not suited to analyzing such relationships. To help identify any non-linear pattern in the way memory performance changes across the lifespan the All-Participants group was split into five equal width age bands, and the band means for key variables plotted. The number of bands (five) was chosen to be the largest possible while still ensuring adequate participants per band to provide meaningful averages. Figure 3.23 illustrates how 55min recall (Fig 3.23a), 24hr recall(Fig 3.23b), learning errors (Fig 3.23c), subjective memory complaints (MCS scores, Fig 3.23d) and 55min-24hr forgetting rate (Fig3.23e) vary by age between 18 and 82 years.

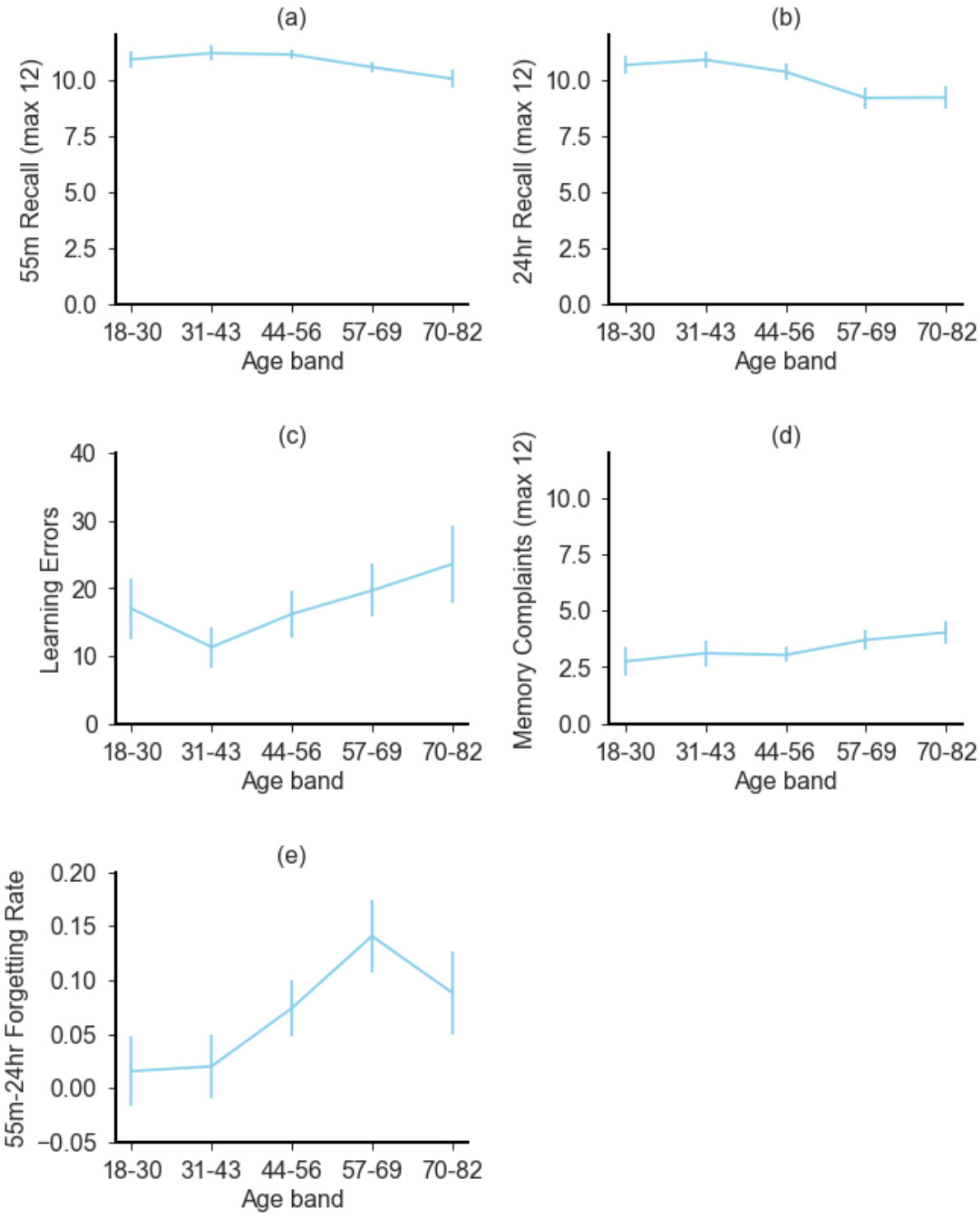


*Figure 3.23 Variation across the lifespan of (a) 55m recall, (b) 24hr recall, (c) learning errors, (d) subjective memory complaints, (e) 55m-24hr forgetting rate (error bars +/- 1SE). Participants per band: 18-30 N=20; 31-43 N=17; 44-56 N=26; 57-69 N=27; 70-82 N=14*

The relatively small scale of the changes between each pair of neighbouring age bands and the number of participants per band means statistical comparisons will not be significant. However, visual inspection indicates a common trend across variables, with performance declining (recall scores decreasing; learning errors, memory complaints and

forgetting rate increasing) by the 57-69 age band and in some cases earlier. More precise identification of the age when deterioration starts will require recruitment of a larger sample. There is a noticeable reduction in the 55m-24hr forgetting rate for the 70-82yr band, compared to the 57-69yr band, which runs against the general trend. This is assumed to be a sampling error due to the small sample size for this band (N = 14).

#### *3.3.3.13 Prevalence of ALF in the general population*

To measure ALF prevalence in the general population, it is first necessary to clarify what this means. The standard definition of ALF requires an acceleration in forgetting in comparison to some control group. For example, most early studies (eg. Butler et al., 2007; McGibbon & Jansari, 2013) compared those with an epilepsy diagnosis (usually TLE or TEA) to a group of control participants with no epilepsy diagnosis. Manes et al. (2008) compared participants with and without an MCI diagnosis, and those reporting and not reporting subjective memory complaints.

However, when referring to prevalence in the general population the comparison group is the entire population, and defining the criteria for ALF becomes an arbitrary statistical exercise. For example, we may decide that we will class the 5% of the population with the fastest forgetting rates as displaying ALF. In that case any discussion of prevalence in the population as a whole is largely irrelevant; by definition the prevalence will be 5%. However, what we can do is analyse the characteristics of this 5%, for example by looking at the ages of people in that group.

Table 3.7 shows a breakdown of those with the highest forgetting rates by age band, using both a 2.5% threshold and a 5% threshold as the criteria for ALF. The actual number observed is shown, and an adjusted number which accounts for the unequal number of participants in each age band in the overall sample. The weighting factor used for each age band was [total participants / (number of age bands * participants in the age band)]. For each age band the weighting factor scales the observed number to what would have been observed if all age bands had an equal number of participants.

*Table 3.7 Frequency of ALF by age band for two possible diagnostic criteria (worst performing 2.5% and 5%).*

| | 2.5% Threshold | | 5% Threshold | |
|---|---|---|---|---|
| **Age band** | **Number observed** | **Adjusted number** | **Number observed** | **Adjusted number** |
| 18-30yrs | 0 | 0 | 0 | 0 |
| 31-43yrs | 0 | 0 | 0 | 0 |
| 44-56yrs | 1 | 0.8 | 3 | 2.4 |
| 57-69yrs | 2 | 1.5 | 2 | 1.5 |
| 70-82yrs | 0 | 0 | 0 | 0 |

Since the overall sample consists of 104 participants, only 3 or 5 will be classified as displaying ALF, based on the 2.5% and 5% thresholds respectively. It appears that ALF is more common in older age groups, first arising in the 44-56yr age group. However, the number of participants classified as displaying ALF is clearly too small to draw any conclusions about prevalence across age bands; any such conclusions will need to wait for future work with a much larger sample size. The proposed method of analysing this is presented here so that it can be followed in any future experiment which has a much larger sample size.

### 3.3.4 Discussion

This experiment used a new online version of the VALMT to investigate memory in a sample from the general population. The aims were: to verify whether the online version produces similar results to the older face-to-face version, to provide concurrent validity; to reduce any possible confounds due to differential interference to see if these may have driven the group differences seen in Experiment 1 and 2; to investigate the relationship between errors made during learning and the subsequent delayed recall scores, particularly for older participants; to investigate forgetting over the 55m to 24hr period and to investigate how memory performance varies across the lifespan. The results will be discussed in that sequence.

#### *3.3.4.1 Comparison with previous results*

Previous experiments using the face-to-face VALMT version tested recall at delays of 5min, 30min and 55min. The key findings from those studies were that by 55min the Older group scored lower than the Younger group, that this difference was largely driven by the slow learning older participants, and that being an Older slow learner correlated

with increased subjective memory complaints. Overall, the online test produced a similar pattern of results at 55min, suggesting both tests are measuring the same thing (concurrent validity) and also strengthening the evidence for these key results.

However, although the same general pattern of results was observed, the size of the differences between Younger and Older groups and between the Older_fast and Older_slow groups was smaller in the online experiment, for learning performance, recall performance, and memory complaints. In some cases the smaller effects resulted in comparisons that had produced statistically significant results with face-to-face testing failing to do so for the new online version. A comparison of age groups across studies showed that while both Younger groups performed very similarly on all measures, the Older group in the online experiment performed better than their equivalents in the face-to-face experiment. This better performance by the online Older group seems to be the cause of the smaller group differences.

One explanation for this is sampling error, as the sample sizes in all VALMT studies so far are relatively small. However, although the samples are small, the sizes are similar between Experiment 2 (face-to-face) and the current experiment. While larger sample sizes would very likely result in a common pattern of statistically significant results across both versions this seems unlikely to be the main driver of the noticeable difference seen in effect sizes.

A second possibility is that a self-selection bias may be present, with those with poorer memories less likely to take part, perhaps for fear of being diagnosed with a memory problem. However, the same argument applies to the recruitment of participants for the face-to-face studies. In addition, to minimise any anxiety the recruitment materials and experiment briefing were very clear that the test would not provide a diagnosis of any sort and that many perfectly healthy people could be expected to score poorly. Furthermore, the fact that the online and face-to-face Older groups are matched on demographic variables, particularly the total memory complaints score, suggests the samples are matched and argues against any differential self-selection bias across studies.

A third possibility is that those with poorer memories may be more likely to drop out without completing all stages or complete their delayed tests at the wrong time. Unlike self-selection at recruitment, this is something that could be objectively assessed. A comparison of the learning performance of those who were excluded for these reasons showed no significant difference from those who completed all stages correctly. This

suggests dropping out or being excluded for completing a delayed test at the wrong delay is also unlikely to have been a major driver of the observed group differences.

A fourth possible explanation is the online test's enforcement of a 20 minute time limit to the learning phase. In a face-to-face test the researcher can observe whether a participant is struggling and intervene if they become stressed, or if it becomes clear that they will be unable to complete learning. In a fully automated online study this intervention is not possible. To prevent the scenario where participants with poor learning performance continue the learning phase indefinitely a 20 minute limit is enforced. If this limit is reached the system tells the participant they have completed the learning phase. The feedback given is exactly the same as that given to those who do complete learning. This means the participant is unaware of their status, reducing any risk of anxiety, and they are allowed to carry on and complete the later stages of the test.

For the comparisons with previous experiments only those who have learnt all 12 pairs to criterion can be included. This means some of the poorest learners will be excluded from the analysis, as they will fail to complete learning within 20mins. In contrast, such participants would have been allowed to continue learning beyond 20mins in the face-to-face version, as long as they did not become stressed, and some may have been able to complete learning and would be included in subsequent analysis. Only one of those who failed to complete learning in the current experiment was aged under 30yrs, while six were aged over 60yrs. Excluding poor learners in this way can therefore be expected to differentially improve the observed learning performance of the Older group, and especially the Older_slow group. Similarly, it would be expected to lead to higher recall scores and reduced memory complaints for these groups.

Analysis of the older excluded poor learners who may have been included if the 20min limit was not imposed confirmed that their mean memory complaints and learning errors were indeed higher than both the combined Older group and the Older_slow group. Including these previously excluded participants back into the analysis for learning performance and memory complaints did, as expected, increase the size of the group differences. In fact, the resulting increase in effect size was sufficient to change the result of statistical comparisons bringing them in-line with the face-to-face results; the variance in the Older groups learning performance becomes statistically greater than the variance of the Younger group, and the correlation between memory complaints and learning performance for the Older group becomes significant. Their inclusion also increased the Older vs Younger group difference in learning trials, moving that test close to

significance (p reduced from .35 to .085), and if these expired participants had been allowed to continue learning (as they were in the face-to-face version) they would have made further errors which may have resulted in this comparison becoming significant too. Considering the strong association between learning errors and subsequent delayed recall it also seems very likely that the resulting group differences in recall would have increased, bringing them more in line with Experiment 2.

Another way to analyse this effect is to simulate a 20min time limit in the face-to-face studies by retrospectively excluding anyone who took longer than 20mins to complete learning during the first learning session. Such analyses for Experiment 2 showed that these exclusions reduced the Younger vs Older group differences as expected, due to the Older group's recall and learning performance both improving. However, in this case the change was not enough to change the result of any statistical tests, and the effect sizes remained large.

Overall, there is strong evidence that the 20min time limit has influenced results by excluding poorer performers in the Older group, and that while this has been a major driver in the smaller observed group differences it is insufficient by itself to account for all of this, and some other factor must be present.

Although the enforced 20min time limit may reduce the group differences seen in a research experiment of this type, it may be beneficial in a clinical setting. The analysis above suggests that the VALMT learning phase may be able to identify those with the greatest cognitive impairment without the need for any delayed recall testing. This would reduce the total effort involved, and enable more rapid diagnosis for these individuals. The delayed recall element could then be used to identify those with a lesser impairment. The combination of a learning measure combined with two delayed tests at increasing intervals may allow the VALMT to identify three categories of impairment: a strong impairment which can be detected during learning, a shorter-term retention problem detected by the first delayed test, and a long-term retention problem identified by the second delayed test at 24hrs or longer. This sequence fits with the stages of progressive memory impairment in Alzheimer's disease proposed by Weston. et al. (2018). In their model, as AD progresses it first impacts long-term retention, then short-term retention, and finally memory encoding. It is possible that by adjusting the delays of the two delayed recall tests in the online VALMT it will be possible to distinguish between these three types of impairment using a single test.

A final possible explanation for the smaller group differences seen in the online version is reduced interference. The online procedure removes the complex interleaving of learning and test used in the face-to-face experiments. It is possible that reduced interference leads to improved recall in the Older, and particularly the Older_Slow, groups. This possibility is discussed in detail below.

#### *3.3.4.2 Impact of reducing possible confounds due to interference*

The detailed analysis of Experiments 1 and 2 identified that interference may influence results in multiple ways. First, within a single learning session pairs may be learnt at different speeds, and for those pairs learnt first the ongoing learning of remaining pairs may provide interference. This 'within learning' interference will be discussed in the next section.

Second, the complex interleaving of learning and testing in the procedure used in the face-to-face experiments may create interference. This may impact observed group differences in delayed recall if groups differ in their sensitivity to interference, for example if Older participants are more vulnerable. This would then be reinforced for the Older_Slow participants who also take more trials to complete learning during each learning session, meaning they would both encounter more interference and be more vulnerable to that interference. To reduce this potential confound the online design uses a greatly simplified procedure, with a single learning session and no other activities interleaved between completion of learning and the delayed test.

Comparing results across studies showed that the Younger groups in the face-to-face and online versions performed similarly for both learning and delayed recall. This suggests that for the 18-30yrs age group any possible interference due to the complex interleaving has negligible effect.

However, for the Older groups large differences were seen, with the Older group in the face-to-face experiment performing worse than the equivalent group in the online experiment for both learning and delayed recall. The fact that they performed worse on both metrics complicates the interpretation of this result. If the learning performance had been similar across experiments then the difference in recall could be attributed to the difference in interference. However, since the learning performance is different, and previous results have shown there is a strong association between learning performance and subsequent delayed recall, it is not clear to what extent the observed difference in

delayed recall is caused by interference, by a difference in the samples themselves, or a combination of these factors.

#### *3.3.4.3 Detailed analysis of learning errors*

One key question when interpreting VALMT results is why some older people make a larger number of errors during learning. Most importantly, does this provide an early indicator of cognitive decline that might be a marker for risk of developing dementia?

A second key question is whether the lower recall performance displayed by some Older participants is caused by making a greater number of errors during learning and the associated interference ('within learning' interference), or whether it reflects a faster underlying forgetting rate, or a combination of these factors. The additional granular statistics gathered by the new online VALMT system make it possible to look at the errors and recall for each individual word-pair, rather than just the total number of errors and total recall figures which the original face-to-face version supplied. This then makes it possible to investigate errors and their impact on recall in greater detail and tease apart the impact of errors and underlying forgetting rate.

With regard to the first question, why some older people make a larger number of errors during learning, one possibility is that those who make more errors are just older, and that it is age that causes the poor performance. However, there was no statistically significant correlation between age and learning errors for either the All_Criteria_Met group or the Older group. This argues against age being a primary driver.

A second possibility is that some older participants are less familiar with using computers and this leads to them making more errors. While computer literacy was not directly measured, time taken per attempt could be used as a proxy. Those who are less familiar would be expected to take longer entering their answer for each trial. While the Older_Slow group did take longer per attempt the difference was small and non-significant, and the change in errors associated such a small change in duration was also small. Together these results suggest that computer familiarity is unlikely to explain the large variation seen in errors made during learning by the Older group.

A third possibility is that those who make most errors are displaying a cognitive decline which may be a pre-clinical stage of AD or other form of dementia. It is known that subjective memory complaints indicate increased risk of dementia (Mitchell et al., 2014; Weston et al., 2018), so if a pre-clinical stage of dementia was the cause of some

older participants higher error rate we might expect those making more errors to also report more memory complaints. While Experiment 2 identified such a significant correlation in the Older group, no significant association was initially found in this experiment. However, this seems to have been at least partly due to exclusion of poor learners who fail to learn all pairs to criterion within the 20min limit, as discussed above (section 3.3.4.1). In fact, including these participants back into the analysis made this correlation significant. There are also theoretical reasons to expect poor VALMT performance to correlate with risk of developing AD. The VALMT uses paired-associate learning, which is vulnerable to the impact of early stage AD (Sapkota et al., 2017) and relies heavily on hippocampal and entorhinal cortex regions which are known to be vulnerable to change in early AD (de Rover et al., 2011; Coupe et al., 2019; Braak & Braak, 1998).

Overall the evidence seems most consistent with increased learning errors in some Older participants being due to some form of cognitive decline, which may indicate risk of developing dementia in the future.

With regards to whether the lower recall performance of some Older participants is caused by interference due to errors or reflects a greater underlying forgetting rate, one interference model suggests that incorrect responses made during learning are retained in some form and create interference for the correct answer. In this case it is specifically the incorrect answers to a given pair that provide interference for that pair and lead to poorer recall. This model suggests that those word-pairs which generate more errors during learning should be recalled less successfully than those which result in fewer errors. An alternative explanation would be that it is the overall total number of errors made during learning that is important. This could be because the process of learning of any given pair provides some interference to the other pairs too, or it could be because it increases the average delay between learning a pair to criterion and eventually recalling it at the 55min test point.

Visual analysis of the number of errors made for each individual word-pair and the subsequent delayed recall showed a clear pattern, with the pairs that encounter most errors being recalled most poorly, which supports an interference explanation. The correlation of these variables (mean number of errors for each pair and the subsequent recall of that pair) confirmed this relationship statistically; the correlation was large and significant. This evidence suggests that the poor delayed recall experienced by some Older participants is at least partly due to interference caused by the larger number of

errors they make during learning. However, perhaps these slow learning older participants also display a greater underlying forgetting rate?

It was impossible to tease apart the impacts of interference and underlying forgetting rate in the studies using the older face-to-face VALMT version due to the limited statistics provided by the software. However, with the newer online version it was possible to look at the errors per-pair, and identify those pairs which were learnt with no errors (*errorless* pairs) and those which generated errors (*errorful* pairs). The statistics for each could then be analysed separately. All errorless pairs for all participants will have encountered exactly the same number of presentations (1 initial presentation plus 3 more when testing learning to criterion), the same number of successful recalls during learning (3), and the same number of learning errors (zero), so learning is very well equated, allowing us to directly compare forgetting rates without the possible confound of differing levels of interference.

Comparison of the forgetting curves for errorful and errorless pairs showed that errorful pairs are forgotten more rapidly, providing further evidence for the interference model linking forgetting to errors during learning.

Between group comparisons of the forgetting curves for errorless pairs found Older participants have a similar underlying forgetting rate to Younger participants between completion of learning to criterion and recall at 55min, and may have a slightly faster forgetting rate between 55min and 24hr. However, importantly, there was no evidence that Older_Slow participants have a faster underlying forgetting rate than Older_Fast participants over either time period. This result held when using the standard relative forgetting formula and the alternative absolute formula. This suggests that the differences in observed recall performance between the Older_Fast and Older_Slow groups do not reflect a difference in their underlying forgetting rate, but must instead be caused by some other factor such as interference.

#### *3.3.4.4 Forgetting between 55min and 24hr*

The 24hr test delay was chosen for the online version to test memory at a longer delay than previous VALMT work, with the aim of identifying any late-onset forgetting that would otherwise have been missed. Comparison of recall across groups found that while the Younger group showed negligible forgetting between 55min and 24hr, the Older group forgot more over this interval, with the Older_Slow group showing the greatest

forgetting, with differences that approached significance. However, since the Older_Slow group also show the lowest recall at 55min, it seems their forgetting has already started before 55min, so their greater forgetting after 55min may not indicate late-onset forgetting, but rather a continuation of forgetting that starts early. When analysis was restricted to those who performed normally for their age group on learning and 55min recall no evidence of late-onset forgetting was found in either the Younger or Older group.

When Weston et al. (2018) found a correlation between ALF and risk of dementia they measured recall at 30min and 7 days, finding no impairment at 30min but a deficit at the longer delay. The loss of retention they detected could be starting anywhere between the two timepoints. It could be that 24hr is not a long enough delay for this type of late-onset forgetting to manifest, in which case switching to a longer delay (e.g. one week) for the second delayed test may be beneficial if VALMT is to capture ALF starting at these longer timescales.

Overall, the 24hr delay does seem to capture some forgetting in the older participants, so is providing valuable information. However, the total forgetting is low, and there may be benefit in pushing this second test out to a longer delay of 3 days or one week to allow more time for differences in forgetting to manifest. This will be of particular benefit if the earliest stages of AD do not impact encoding or early retention, as Weston et al. (2018) suggest, and therefore cannot be detected by looking at learning performance or recall at 55min. Alternatively, if encoding and/or early retention are always impacted then it may be that the more sensitive encoding test provided by the VALMT allows diagnosis without the need for any testing at such extended delays. Deciding between these possibilities will require further work with longer delays, and preferably with those known to be in the early stages of AD.

Although memory complaints correlated with recall at 55min, as seen in previous experiments, by 24hrs the correlation had reduced and become non-significant. While this could mean that people are more aware of their memory performance at short delays than long delays, it seems more likely that the 24hr delays allows more time for other factors to add variance which then weakens the correlation.

#### *3.3.4.5 Memory performance across the lifespan*

This experiment recruited adult participants of all ages with the aim of assessing how memory develops across the lifespan, in particular memory measured at extended delays beyond the 30min typical in memory research. For example, is it stable until old age, or does it perhaps start to decline in middle age?

Initial analysis across all participants found small but significant negative correlations between age and delayed recall at both 55min and 24hr. However, while this tells us that older people recall fewer items, it does not help answer questions about when deterioration starts, and whether this continues at a constant rate or accelerates. Indeed, since correlation measures linear relationships a lack of a significant correlation only identifies the lack of a linear relationship, not the lack of *any* relationship; it is still possible that a non-linear relationship exists. To check for such non-linear patterns participants were split into five equal width age bands, and the band means for key variables plotted. Visual inspection identified a common trend, with performance declining (recall scores decreasing; learning errors, memory complaints and forgetting rate increasing) by the 57-69 age band and in some cases earlier. This suggests decline in memory performance starts in late middle-age, however more precise identification of the age when deterioration starts will require recruitment of a larger sample. In the only previous lifespan analysis of forgetting rates over extended delays Davis et al. (2003) found evidence of accelerated forgetting in the oldest age group (76-90yr). The data from the current experiment indicates such forgetting may start at an earlier age, but again, a larger sample size will be required to statistically validate this.

### 3.3.5 Conclusion

This experiment is the first to use the new online VALMT with participants across a wide age range. The results followed the same pattern as the earlier face-to-face studies, confirming that Older participants score lower on delayed recall by 55mins, that this difference is driven by slower learning older participants, and that being a slow learner is associated with increased memory complaints. However, the size of the group differences observed were smaller than in face-to-face studies, and in some cases not statistically significant. Evidence suggests this was partly driven by excluding those who failed to complete learning within 20mins, which eliminates the poorest performers. Another possible cause is the elimination of the complex interleaving of learning and testing,

which reduces opportunities for interference to impact results. Analysis suggests that while this does not impact the Younger age-group, it may have impacted the Older age-group, although it is not clear to what extent the reduced group differences in the online version are caused by the removal of such interference, or by a difference in the learning performance of the Older samples across studies.

Analysis of forgetting for errorless pairs indicated no difference in underlying forgetting rate between the fast and slow learning older participants. Instead, there was clear evidence that the strongest influence on delayed recall is the number of errors made during learning. The evidence also indicates that the most likely cause of the increased learning errors made by some Older participants is some form of cognitive decline, which may indicate risk of developing dementia in future. Considering these results in combination, the most likely explanation of the differences in delayed recall between fast and slow learning older participants is that the slow learners have a learning deficit, reflecting cognitive decline, which causes them to make more errors during the learning process, that these errors create interference for the specific pair being learnt, and that this in turn leads to poorer recall.

The 24hr delay captured some forgetting in the Older group; however, once analysis was restricted to those who perform normally on learning and 55min recall there was no evidence of any late-onset forgetting in any group. The value of this extended delay for diagnosing those at risk of AD will depend on whether the impact of the earliest stages of the disease is restricted to triggering late-onset forgetting (in which case the extended delay will be needed, as learning and recall at 55min will not be impacted), or whether it also impacts learning and early retention (in which case the extended delay is not essential and measuring learning and recall at 55min alone may be adequate); further work will be required to tease apart these possibilities.

Analysis of memory across the lifespan showed a trend for performance to have declined (recall scores decreasing; learning errors, memory complaints and forgetting rate increasing) by the 57-69 age band; however, a large sample will be required to more precisely identify when this deterioration starts. A larger sample size will also be required to allow the relative prevalence of ALF by age group to be analysed.

## 3.4 Experiment 5 - Online VALMT testing with 16-17yr age range

### 3.4.1 Rationale

This is one of a very few experiments of its kind, looking at memory and forgetting over extended delays in the younger 16-17yrs age group; a search of the literature found only two such previous studies. This experiment had two primary aims. First, it aimed to test the online VALMT procedure with this younger age group, and highlight any issues with the process. Second, it aimed to compare results for this younger group with those for other age ranges, extending the lifespan analysis from Experiment 4 to include all ages down to 16yrs old.

### 3.4.2 Methods

#### *3.4.2.1 Participants*

##### 3.4.2.1.1 Recruitment

Participants were recruited by advertising the study to UK secondary school psychology teachers via social media. Teachers were asked to involve students from the last two years of secondary education, which typically covers the ages from 16 to 18 inclusive. For a planned comparison with the 18-30yr group from Experiment 4, the relatively small sample size for that group (N=20) means achieving a power of 80% for detection of anything other than a large effect size would be unrealistic. To detect a large effect size a sample size of 20 was required.

##### 3.4.2.1.2 Inclusion criteria

To be included in the main analyses a participant had to meet the same requirements as Experiments 3 & 4, except that an age range criterion was added. The criteria and the number failing to meet each (number in [brackets]) are summarised below. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as some excluded participants people failed multiple criteria:

1. Must not report dyslexia [8]
2. Must not report a medical condition that might impact memory [1]
3. Must be aged between 16 and 17yrs, inclusive [7]
4. Reported English language level must be either first language or fluent [2]

5. Must learn all 12 pairs to criterion within 20 minutes [2]
6. Must complete all 3 stages: learning, 55min & 24hr tests [30]
7. Must complete the 55min test between 45 and 65min (55min +/- 10min) [43]
8. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [0]

Six participants were aged 18, and one was 19yrs old. This is due to variation in the age range of students in the final year of secondary education in the UK. To avoid overlapping age ranges with previous studies which had 18-30yr age groups, these 18 and 19yr old participants were excluded from the main analyses.

Participants were instructed to complete the delayed tests as close to the requested time as possible (55mins, 24hrs), but to still complete the delayed tests even if they were unable to come back at the correct time. The timing data showed that some teachers had reduced the delay for the first delayed test to allow them to run the learning stage and the first delayed test within a single lesson. This meant many students completed their first delayed test early, between 25min and 45min. To maximise the use of data three groups were prepared. First, an *All_Criteria_met* group containing only those who met all criteria, which could be confidently compared with data from previous studies. Second, an *Early_Completer* group who met all criteria except that they completed their first delayed test early, in the 25-45min range. Finally, a *Combined* group containing all participants from both the All_Criteria_Met and Early_Completer groups.

As in previous experiments, the risk of distortion of results due to excluding those who dropped out or failed to complete delayed tests at the correct time was analysed. A comparison was performed between the All_Criteria_Met group who met all criteria and those who met criteria 1, 2 ,3 and 4 but failed to meet any of criteria 5 ,6, 7 or 8. This analysis showed that the two groups were not significantly different on errors made during learning, which suggests excluding these participants will not impact the results ($Errors_{excluded}$ = 16.6, $Errors_{included}$ = 17.92, MWU = 259, p=.78, r = 0.06, BF = 0.32).

#### 3.4.2.1.3 Included participants

A total of 78 individuals took part. The demographics of the three groups created after applying exclusion criteria are summarised in Table 3.8.

*Table 3.8 Demographic information as a function of group*

| Factor | All_Criteria_Met | Early_Completer | Combined |
|---|---|---|---|
| N | 12 | 9 | 21 |
| Gender | 2M/10F | 2M/7F | 4M/17F |
| Age Mean(SD) | 16.75 (0.45) | 16.33 (0.50) | 16.57 (0.51) |
| Language: | | | |
| First Language | 11 | 8 | 19 |
| Fluent | 1 | 1 | 2 |

As all participants are at the same stage of their education there was no need to track an education variable. The All_Criteria_Met and Early_Completer groups were matched on Gender ($X^2(1) = 0.10$, $p = .75$, BF = 0.62), Age (t(19) = 1.99, p = .06, Cohen's d = 0.89, BF = 1.49) and Language ($X^2(1) = 0.05$, $p = .83$, BF = 0.71).

**NOTE:** The resulting group sizes after applying all exclusion criteria are small. The following analysis should therefore be treated as provisional; more data will be required before confident conclusions can be drawn.

#### *3.4.2.2 Stimuli*

The stimuli set consisted of the same 12 unrelated word-pairs used in the previous experiments (refer to Experiment 3 for detail).

#### *3.4.2.3 Procedure*

The new standard online VALMT procedure was used; refer to Experiment 3 for detail.

### 3.4.3 Results

#### *3.4.3.1 Learning performance:*

If the Early_Completer group took their first delayed test early because of the way their teacher coordinated activity during a lesson, rather than due to anything about the participant themselves, then we would not expect any association between learning performance and being an Early_Completer. To investigate this Table 3.9 compares the

learning performance of those who met all criteria with those who met all criteria except for completing their first delayed test too early.

*Table 3.9 Trials to learn to criterion as a function of group*

| Factor | All_Criteria_Met (N=12) *Mean(SD)* | Early_Completer (N=9) *Mean(SD)* | Combined (N=21) *Mean(SD)* |
|---|---|---|---|
| Trials | 53.91(14.64) | 49.33(12.73) | 51.95(13.71) |

There was no significant difference between the learning trials needed to reach criterion for the All_Criteria_Met and Early_Completer groups, ($Mdn_{All\_Criteria\_Met}$ = 56.0, $Mdn_{Early\_Completer}$ = 48.0, *MWU* = 63.5, $p$ = .52, $r$ = 0.18, BF = 0.45). Although the sample sizes are small, this suggests these two groups are similar on this variable, and that it is acceptable to use the aggregate Combined group for some later analyses of learning performance, making better use of available data.

### *3.4.3.2 VALMT delayed cued-recall performance*

To investigate delayed recall performance and forgetting rates the mean recall scores at 55min and 24hr are plotted in Figure 3.24, for the All_Criteria_Met group. To illustrate the impact of early completion of the 55m test the data for the Early_Completer group is also shown.

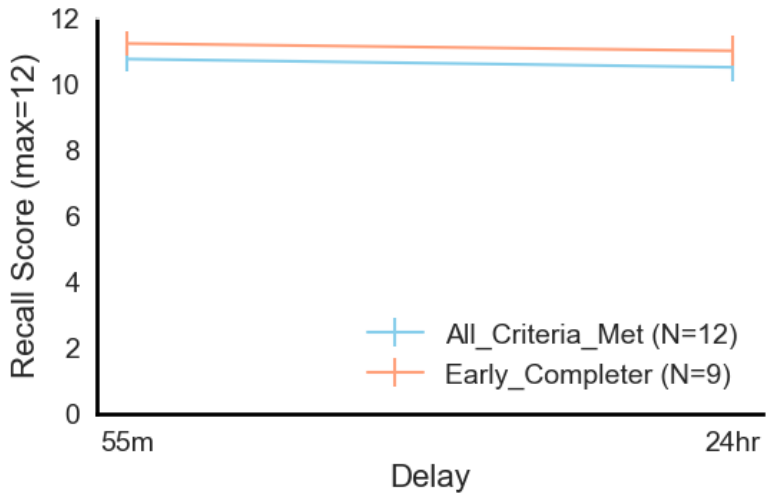


*Figure 3.24 Delayed recall performance at 55min and 24hr delays for the All_Criteria_Met and Early_Completer groups (error bars +/- 1SE).*

The variation in cued recall performance between groups and across delay intervals was analysed using a mixed factors ANOVA with within-subjects factor Delay (55min vs

24hr) and between-subjects factor Group (All_Criteria_Met vs Early_Completer). There was no significant main effect of Group ($F(1, 19) = 0.81$, $p = .38$, $\eta_p^2 = 0.04$, BF = 0.52) or Delay ($F(1, 19) = 1.38$, $p = .26$, $\eta_p^2 = 0.01$, BF = 0.42) and no significant interaction ($F(1, 19) = .01$, $p = .95$, $\eta_p^2 < 0.001$, BF = 0.22). This indicates that compared to the All_Criteria_Met group the Early_Completer group had a very similar level of recall (Marginal means: $M_{\text{Older}} = 9.72$ pairs, $M_{\text{Younger}} = 10.77$ pairs), that there was very little forgetting between 55m and 24hr, and no difference in forgetting rates between groups.

This analysis suggests the early completion of the first delayed test has limited impact on the recalls scores at either delay. However, although the group difference is small, the Early_Completer group does score higher, and the effect size suggests that with a larger sample the difference would be significant. Therefore, to avoid distorting analyses only the All_Criteria_Met group will used for some of the later analyses and for cross-study comparisons.

#### *3.4.3.3 Distribution of learning errors across word-pairs and their relationship to recall*

To evaluate the effectiveness of each word-pair for this younger age group, and the relationship between learning errors and subsequent recall, Figure 3.25 illustrates the mean number of errors made for each individual pair and the corresponding mean recall rate at each delay. Since this analysis is interested in the general relationships rather group differences it is valid to use the larger Combined group which includes the Early_Completer group; the data shown is for this larger group.

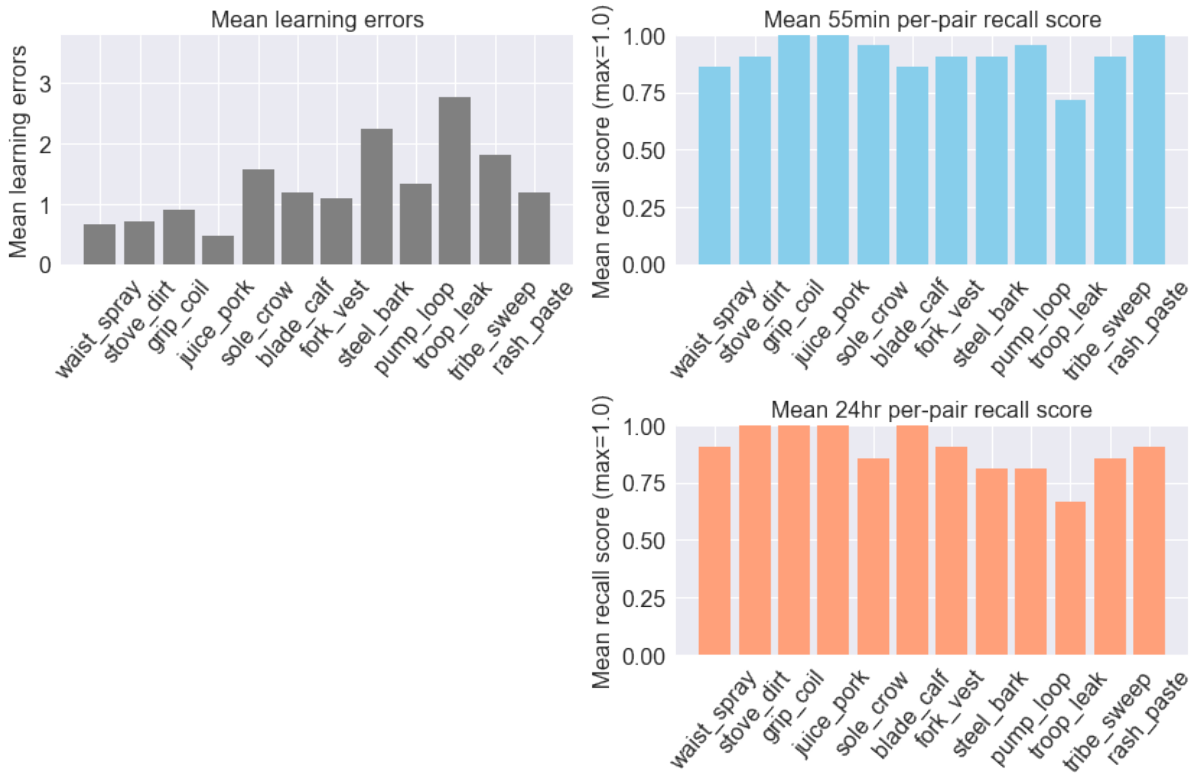


*Figure 3.25 Mean learning errors and delayed recall by individual word-pair for the Combined group (N=21).*

The data shows a similar pattern of results as found in Experiment 3 and 4: there is variation in difficulty between pairs which is beneficial for distinguishing between high and low performers; no pairs encounter zero or a very high number of errors; pairs which encounter most errors are recalled more poorly at both delays, while those pairs which encounter fewest errors are recalled best. However, unlike earlier experiments a small number of pairs were recalled by all participants, so mean recall for these is 100%. This means these pairs are not helping to distinguish between the performance levels of these healthy 16-17yr olds. There is variation in which pairs are recalled perfectly at 55min and at 24hr, which suggests sampling effects. Therefore, with a larger sample size it may be that no pairs would be perfectly recalled by all participants.

#### *3.4.3.4 Relationship between delayed recall and total and per-pair learning errors*

To investigate the relationship between errors made during learning and subsequent delayed recall, the scatterplots in Figure 3.26 illustrate the association between the total number of errors made and the cued recall scores at each delay. Since this analysis is interested in the general relationships rather group differences it is valid to use the larger

Combined group which includes the Early_Completer group; the data shown is for this larger group.

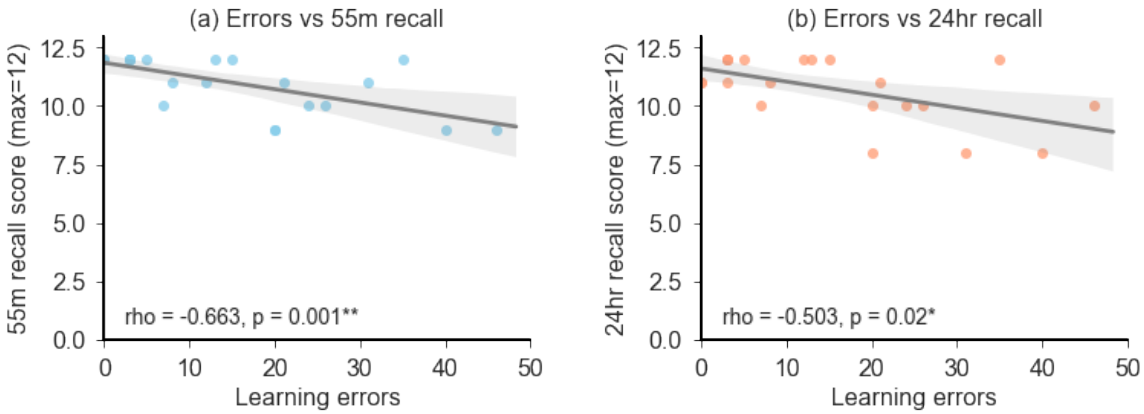


*Figure 3.26 Correlation of total learning errors with delayed recall for the Combined group (N=21); shaded area is 95% confidence interval.*

There was a large and statistically significant negative correlation between learning errors and delayed recall at both 55min and 24hr (55min rho = -0.67, p < .001; 24hr rho = -0.50, p = .02). This indicates that those who make the most errors subsequently recall the fewest pairs, replicating the pattern seen in previous studies.

To investigate this association at the more granular per-pair level, the scatterplots in Figure 3.27 illustrate the relationship between the mean number of errors made learning each individual pair and the mean recall scores for that specific pair, for the Combined group. These plots show one data-point for each of the 12 word-pairs.

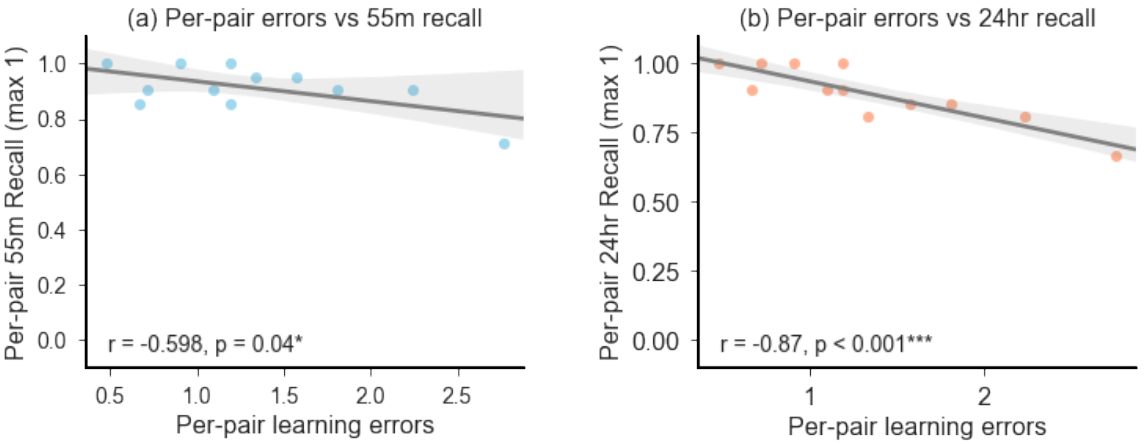


*Figure 3.27 Correlation of learning errors per word-pair with delayed recall per word-pair at (a) 55m and (b) 24hr for the Combined group (N=21); shaded area is 95% confidence interval.*

There was a large and significant negative correlation between learning errors per individual word-pair and subsequent delayed recall at both 55min and 24hr (55min r = -

0.60, p = .04, BF = 2.36; 24hr r = -0.87, p < .001, BF = 138). Overall, this analysis confirms the visual pattern seen in Figure 3.25, in which those pairs which encounter most errors are recalled most poorly.

#### *3.4.3.5 Memory performance across the lifespan*

To investigate how the performance of the 16-17yr group fits into the development of memory with age, the lifespan analysis from Experiment 4 was extended to include this group as a separate age band, and is illustrated in Figure 3.28. Note that the other age bands in the analysis cover 13 years each, while the 16-17yr band covers only 2 years, so the inclusion of the narrow 16-17yr band in this figure should be interpreted with care.

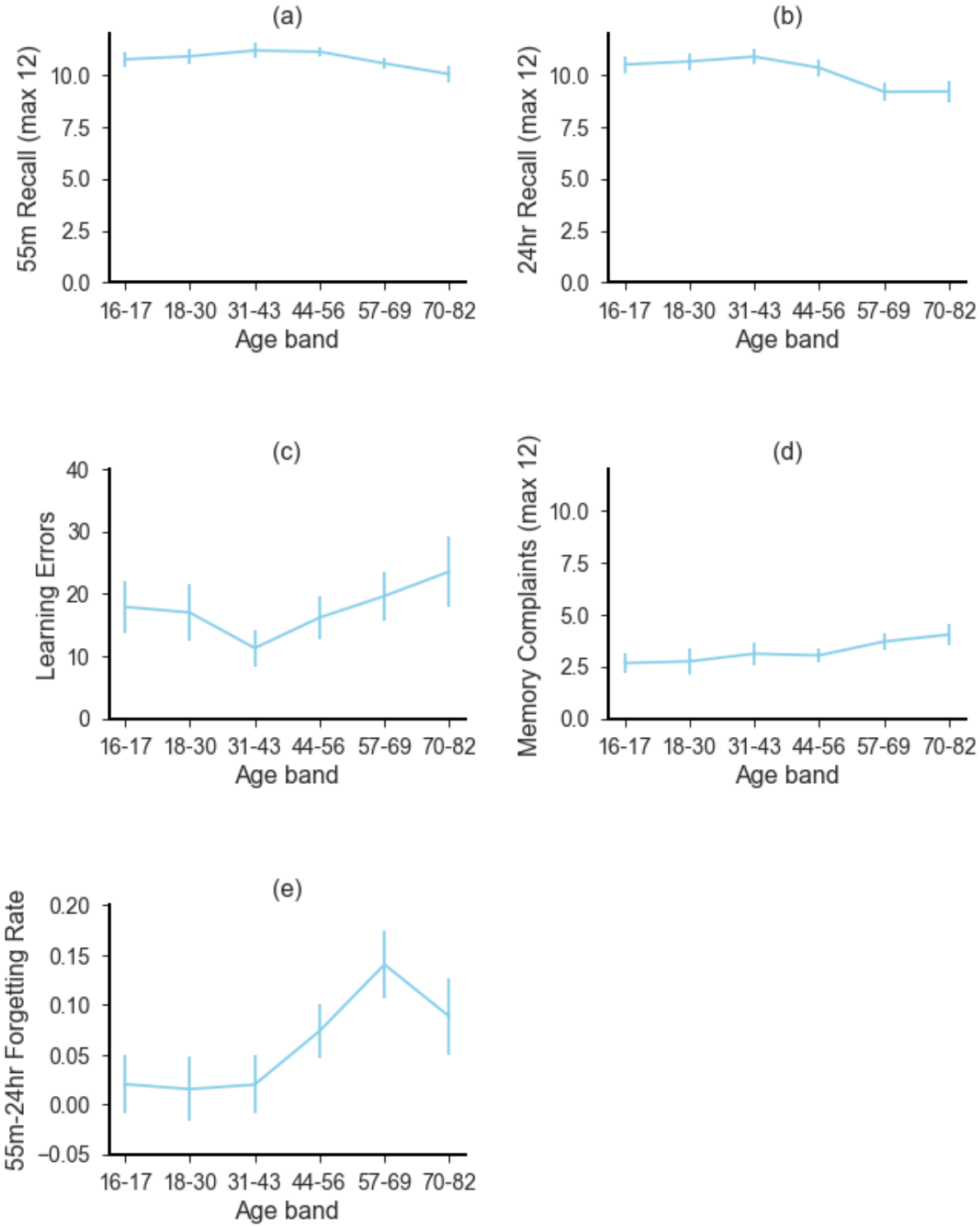


*Figure 3.28 Variation across the lifespan of (a) 55m recall, (b) 24hr recall, (c) learning errors, (d) subjective memory complaints, (e) 55m-24hr forgetting rate (error bars +/- 1SE). Participants per band: 16-17yr N=12; 18-30 N=20; 31-43 N=17; 44-56 N=26; 57-69 N=27; 70-82 N=14*

To further investigate how this 16-17yr old group compares to other ages, Table 3.10 compares key variables for this experiment with equivalent values from the 18-30yr group from Experiment 4.

*Table 3.10 Comparison of key VALMT variables for the All_Criteria_Met group and the 18-30yrs group from Experiment 4.*

| Factor | Expt 4 18-30yrs *Mean(SD)* | Expt 5 16-17yrs All_Criteria_Met *Mean(SD)* | Statistical Comparison |
|---|---|---|---|
| Trials to criterion | Mdn 44.5 IQR 19.75 | Mdn 56.0 IQR 21.25 | $MWU(30) = 134$, $p = .56$, $r = 0.12$, BF = 0.38 |
| 55m recall | 10.90(1.61) | 10.75(1.23) | $t(30) = 0.27$, $p = .79$, $d = 0.10$, BF = 0.35 |
| 24hr recall | 10.65(1.74) | 10.50(1.38) | $t(30) = 0.25$, $p = .81$, $d = 0.09$, BF = 0.35 |
| MCS total | 2.75(2.90) | 2.67(1.72) | $t(30) = 0.09$, $p = .93$, $d = 0.03$, BF = 0.34 |

Participants for the current experiment were deliberately selected to be younger and at a different stage of their education (still at school) so, as expected, they were statistically different on both these variables (Age: t(30) = 4.26, p < .001, Cohen's d = 1.56, BF = 123; Education: $X^2(7) = 11.65$, $p = .009$, BF = 24.61). The groups were, however, matched on Gender ($X^2(1) = 0.05$, $p = .82$, BF = 0.49) and Language ($X^2(1) = 1.37$, $p = .24$, BF = 0.85), suggesting cross-study comparisons are valid.

The trials to criterion, total memory complaints score and delayed recall scores at both delays are very similar for both groups. As a result, statistical tests found no significant differences between the groups on any of these variables (see Table 3.10 for statistics). These results suggest that compared to those in the 18-30yr range individuals in the 16-17yr age range report a similar number of memory complaints, learn at a similar speed and forget at a similar rate over the first 24 hours.

To further illustrate the effect of age on forgetting, Figure 3.29 compares the delayed recall scores at 55min and 24hr for each age band.

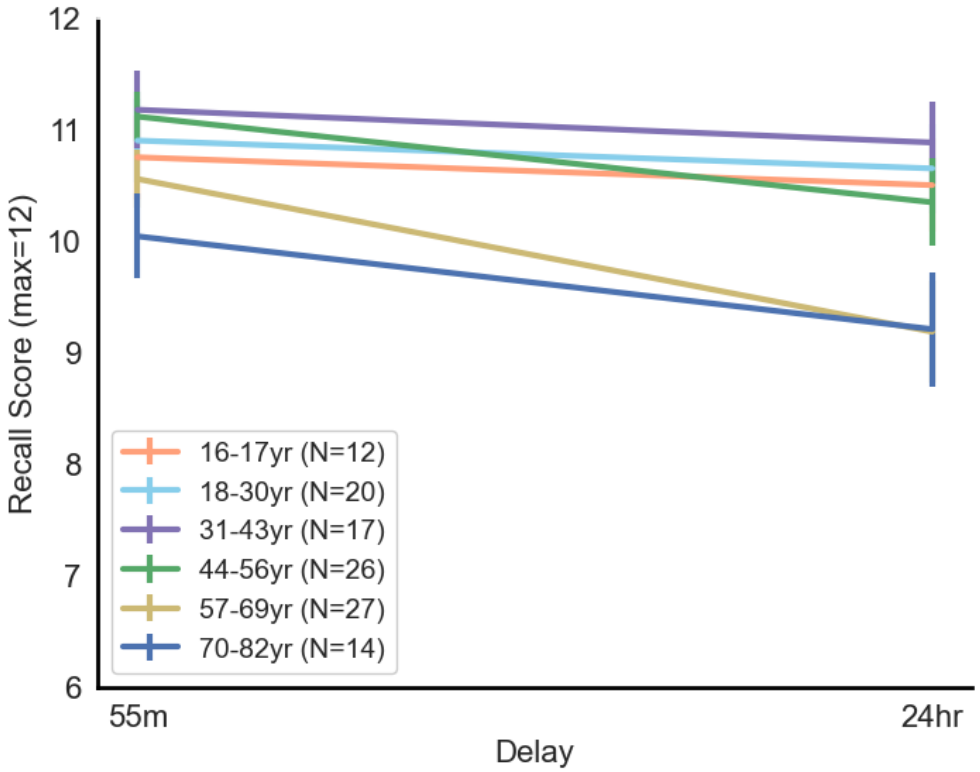


*Figure 3.29 Variation across the lifespan of delayed recall performance at 55min and 24hr (error bars +/- 1SE).*

The 3 youngest age bands perform very similarly, both in recall scores and forgetting rates. Although the small sample sizes mean the comparisons are not statistically significant, visual inspection shows a trend for recall scores to decline from the 57-69yr group onwards (yellow line). There is also a trend for the forgetting rate (slope of the forgetting curves) to increase from the 44-56yr band (green line) onwards. However, as noted previously in section 3.3.3.12, the observed forgetting rate slows again in the 70-82yr age band (dark blue line). One possible explanation for this pattern is that it reflects a sampling error, due to the small sample size in this age band (N=14). An alternative explanation is that the lower baseline scoring in this group may in some way reduce the amount of forgetting observed.

### 3.4.4 Discussion

This is one of a very few experiments of its kind, investigating memory and forgetting over extended delays in the younger 16-17yr age group. The experiment succeeded in validating the online VALMT procedure for use with this age group. Sample sizes for some analyses were small (N = 21 or 12 depending on the variable being analysed),

which mean some conclusions about VALMT performance are provisional. However, bearing this in mind, the results show how the 16-17yr age group compare to older age ranges, fitting them into the wider lifespan analysis started in Experiment 4.

#### *3.4.4.1 Online testing procedure*

A total of 78 participants in the 16-17yr age range took part, and none reported any issues with the online process or technical issues with the test, despite accessing it from a range of devices and operating systems. It appears some teachers ran the 55min delayed test early, to allow it to be completed within a single lesson. While this means those participants' data cannot be used for some of the analyses, this does not reflect a weakness in the test itself. Overall, the online system appears to work well for this age range.

#### *3.4.4.2 Analysis of learning errors and implications for the role of interference*

Analysis of the total number of errors made during learning, the distribution of errors across the 12 word-pairs and the impact of errors on subsequent delayed recall showed the same pattern observed in other experiments; those who make the most errors score lowest at delayed recall at both delays, and those pairs which generate the most errors are recalled most poorly. This suggests the same causal pattern identified in Experiment 4 is at play, with errors made during learning of a given pair disrupting later recall of that specific pair.

It is expected that such young participants would not have any age related cognitive deficits, so the fact that they show the same link between errors and delayed recall suggests this mechanism is common across all ages. This in turn suggests that the poor recall displayed by some older participants is caused by cognitive change which causes them to make more errors during learning, rather than ageing leading to the emergence of the mechanism linking errors to poor recall. This would fit the model proposed in Experiment 4, in which early stage AD or another dementia process impacts learning, leading to errors, and those errors then lead to poorer recall.

Unlike previous experiments, it was noted that some pairs were recalled correctly by all participants. While this likely to be due to the small sample size, it indicates some pairs may not be adding value for this young age group.

#### *3.4.4.3 Memory performance across the lifespan*

The 16-17yr group performed very similarly to the 18-30yr group from Experiment 4 on all metrics. There was very little difference for trials required to reach criterion, memory complaints score, forgetting rate or delayed recall scores at both delays, suggesting that both age ranges learn at a similar speed and forget at a similar rate over the first 24 hours. It would generally be expected that cognitive abilities such as memory performance as measured by the VALMT would be lower at very young ages while significant brain development is still occurring, then rise to a peak in early adulthood. The results from this experiment suggest that by 16-17 years of age memory performance is already getting close to the peak level.

### 3.4.5 Conclusion

This experiment validated use of the online VALMT with younger participants, in the 16-17yr age range. It also showed that participants of this age are very similar to those in the 18-30yr age range in terms of learning and memory performance as measured by the VALMT, although the small sample size means additional data is required to strengthen these conclusions.

## 3.5 Experiment 6 - Testing domain specificity of VALMT

### 3.5.1 Rationale

The VALMT is intended to measure verbal memory performance. However, it is an open question to what extent VALMT performance reflects pure verbal ability, and how much it reflects some more general memory or cognitive ability. One way to examine this is to compare VALMT performance with a measurement from another cognitive domain. This experiment takes this approach, and uses face recognition memory as the alternate measure. The aim is to investigate whether group differences in memory performance detected by VALMT are specific to the verbal domain or correlate with this non-verbal form of memory.

### 3.5.2 Methods

#### *3.5.2.1 Participants*

##### 3.5.2.1.1 Recruitment

Participants were recruited from a University of Greenwich database. This database holds facial recognition performance data for a large number of people of all ages with differing facial recognition abilities. Two groups of participants were identified. First, a set of high performing participants from those who are classed as 'Super-recognisers'. Second, a set of participants with typical facial recognition abilities.

The exact criteria for selection were those used in previously published face recognition research (e.g. Noyes et al., 2021). To be classed as a super-recogniser a participant must have achieved the maximum score (40 out of 40; achievable by approximately 2.5% of the population) on the Glasgow face matching test (GFMT; Burton et al., 2010), and score in the top 2% of a representative UK sample ($n$ = 254, $M$ = 70.7, $SD$ = 12.3) on the Cambridge face matching test (CFMT+; Bobak et al., 2016). To be classed as typical-ability, participants must score within 1 SD of the mean of Burton et al.'s (2010: $n$ = 192, $M$ = 32.5, $SD$ = 9.7) and Bobak et al.'s (2016) GFMT and CFMT+ samples, respectively.

The two groups will be referred to as *Super-recogniser* and *Typical*. Both sets were approached by direct email advertising the study. A total of 6772 potential participants were contacted, of which 495 started the experiment.

As there was no previous research to indicate an expected effect size the experiment was designed to detect at least a medium effect (Cohen's d = 0.5). Statistical power analysis indicated the group comparisons required sample sizes of at least 51 in each group to provide power of greater than 80%.

##### 3.5.2.1.2 Inclusion criteria

To be included in the main analyses a participant had to meet the following requirements, where the number in [brackets] is the number who failed to meet that criteria. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as many excluded participants people failed multiple criteria:

1. Must not report dyslexia [15]
2. Must not report a medical condition that might impact memory [21]

3. Reported English language level must be either first language or fluent [31]
4. Must learn all 12 pairs to criterion within 20 minutes [19]
5. Must complete all 3 stages: learning, 55min & 24hr tests [47]
6. Must complete the 55min test between 45 and 65min (55min +/- 10min) [59]
7. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [37]

The risk of distortion of results due to excluding those who dropped out or failed to complete delayed tests at the correct time was analysed. A group comparison was performed between those who met all criteria and those who met criteria 1,2,3 & 4 but failed to meet any of the criteria 5 ,6 & 7. The two groups were compared on the number of errors made during learning. Unlike previous studies, this analysis showed that while the effect size was small, the difference between the groups was statistically significant, with the excluded group making more errors ($Errors_{included}$ = 5.0, $Errors_{excluded}$ = 13.0, MWU = 8281, $p < .001$, r = .26, BF = 16.92 ), which suggests excluding these participants has the potential to impact the results by inflating the performance of the included group. This may also impact comparisons of Super-recogniser and Typical groups if the effect of the exclusions is different for these two groups. To investigate this excluded and included participants were compared separately for each group. For both groups it was the excluded participants who made most errors, and the effect size was similar (Super-recogniser: $Errors_{included}$ = 3.0, $Errors_{excluded}$ = 5.0, MWU = 610, p = .22, r = .19, BF = 0.48; Typical: $Errors_{included}$ = 8.0, $Errors_{excluded}$ = 14.0, MWU = 3946, p = .02, r = .21, BF = 1.78) ), which suggests that although the exclusions may inflate the performance of the included groups, it is unlikely to significantly distort comparisons between Super-recogniser and Typical groups.

#### 3.5.2.1.3 Included participants

The demographics of those participants who met all inclusion criteria are summarised in Table 3.11, broken into Super-recogniser and Typical groups.

*Table 3.11 Demographic information as a function of group*

| Factor | Typical | Super-recogniser |
|---|---|---|
| N(Gender) | 85(27M/58F) | 49(5M/44F) |
| Age Mean(SD) | 47.67(14.3) | 40.10(7.94) |
| Education: | | |
| Batchelors | 35 | 23 |
| Diploma | 4 | 5 |
| Doctorate | 4 | 2 |
| High School | 8 | 6 |
| Masters | 20 | 5 |
| Secondary | 10 | 5 |
| Technical | 4 | 3 |
| Language: | | |
| First Language | 54 | 36 |
| Fluent | 31 | 13 |
| MCS Total Mean(SD) | 2.37(1.80) | 2.61(2.15) |

While groups were matched on Education ($X^2(6) = 5.05$, $p = .54$, BF = 0.03), Language ($X^2(1) = 1.39$, $p = .24$, BF = 0.43) and MCS total ((t(132) = 0.71, p = .48, d = 0.13, BF = 0.24), there was a significant difference on Age and Gender, with the Super-recogniser group being younger than the Typical group on average (t(132) = 3.40, p < .001, d = 0.61, BF = 31.72) and containing fewer males ($X^2(1) = 7.95$, $p = .005$, BF = 14.90).

### *3.5.2.2 Stimuli*

The stimuli set consisted of the same 12 unrelated word-pairs used in previous experiments. Refer to Experiment 3 for detail.

### *3.5.2.3 Procedure*

The entire procedure, including gathering of demographics, providing consent, learning and testing was performed online using the new online VALMT. Refer to Experiment 3 for details.

### 3.5.3 Results

#### *3.5.3.1 Learning performance:*

To investigate variation in learning performance Figure 3.30 shows the distribution of errors made during learning for each group.

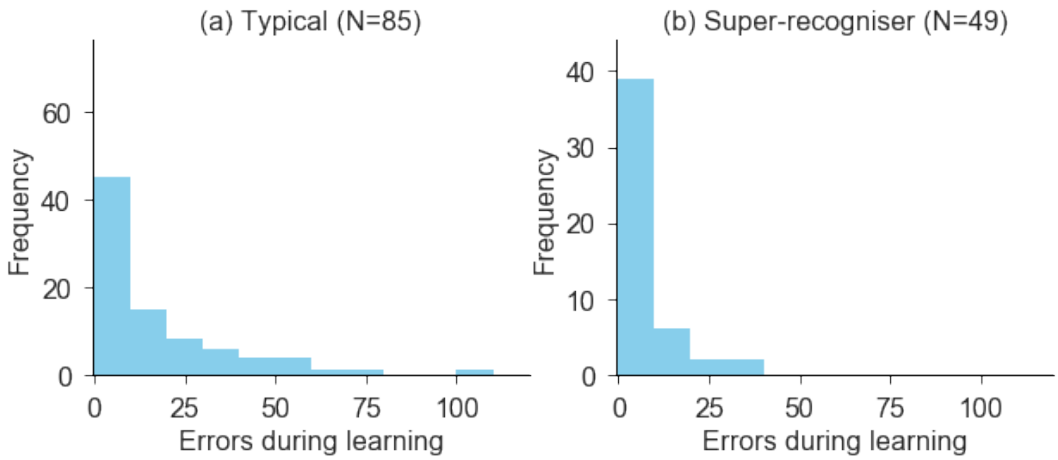


*Figure 3.30 Distribution of total errors made during learning to criterion as a function of group.*

Visual inspection indicates that the distribution for the Super-recogniser group is narrower, with a greater skew towards a lower number of errors. A statistical comparison confirmed that Super-recognisers make fewer errors during learning ($Errors_{Typical}$ = 8.0, $Errors_{Super\text{-}recogniser}$ = 3.0, MWU = 1371, $p < .001$, $r = .34$, BF = 9.70).

#### *3.5.3.2 VALMT delayed cued-recall performance*

To investigate delayed recall performance and forgetting the mean recall scores for each group are plotted in Figure 3.31.

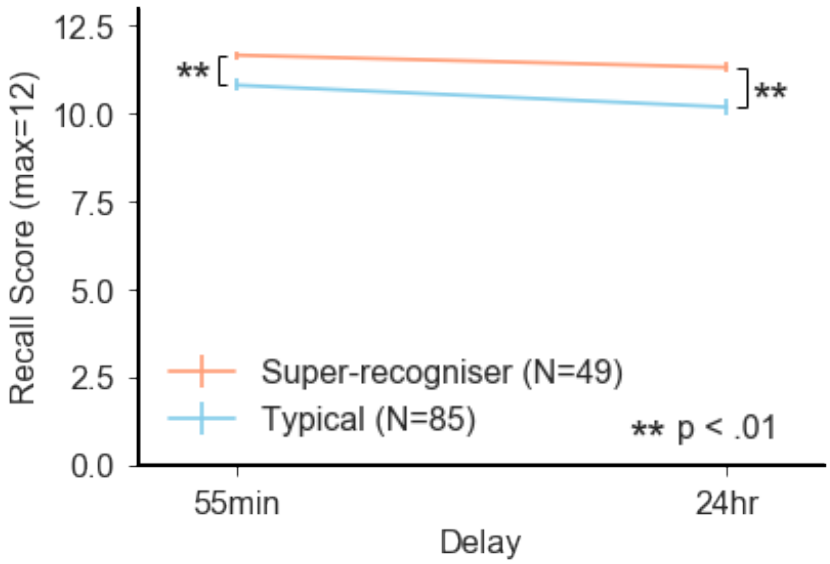


*Figure 3.31 Delayed recall performance at 55min and 24hr delays as a function of group (error bars +/- 1SE).*

The variation in cued recall performance between groups and across delay intervals was analysed using a mixed factors ANOVA with within-subjects factor Delay (55min vs 24hr) and between-subjects factor Group (Super-recogniser vs Typical). There were significant main effects of Group ($F(1, 132) = 13.69$, $p < .001$, $\eta_p^2 = 0.09$, BF = 72.81) and Delay ($F(1, 132) = 15.03$, $p < .001$, $\eta_p^2 = 0.10$, BF = 489) but no significant interaction ($F(1, 132) = 1.22$, $p = .27$, $\eta_p^2 = 0.009$, BF = 1.43). This indicates that compared to the Typical group the Super-recogniser group had an overall higher level of recall (Marginal means: $M_{Older} = 9.72$ pairs, $M_{Younger} = 10.77$ pairs), and while there is a significant amount of forgetting between 55min and 24hrs overall, there was no difference in forgetting rates between groups.

To compare cued recall performance across groups at each time-point two independent sample t-tests were used. The Super-recogniser group scored significantly higher than the Typical group at both 55min ($M_{Typical} = 10.80$, $M_{Super\text{-}recogniser} = 11.65$, $t(132) = 3.38$, Bonferroni adjusted $p = .001$, $d = 0.61$, BF = 30.06) and 24hr ($M_{Typical} = 10.18$, $M_{Super\text{-}recogniser} = 11.31$, $t(132) = 3.39$, Bonferroni adjusted $p = .001$, $d = 0.61$, BF = 30.19). These group differences confirm a significant relationship between VALMT scores and face recognition ability, with those with superior face recognition ability (Super-recognisers) also scoring higher on VALMT at both delays.

### 3.5.3.3 *Subjective memory complaints*

To investigate group differences in memory complaints the mean MCS scores for the Typical and Super-recogniser groups were compared using an independent sample t-test. The group difference was very small, and non-significant ($M_{Typical}$ = 2.36, $M_{Super\text{-}recogniser}$ = 2.61 $t(132) = 0.71$, $p = .48$, $d = 0.13$, BF = 0.24). This suggests subjective memory complaints are not related to the group differences seen in learning performance and recall in these groups

### 3.5.3.4 *Relationship between age and VALMT variables*

Since there was a significance difference in the mean age of the Super-recogniser and Typical groups, with the Super-recogniser group being younger on average, it is important to understand whether this age difference might have driven the observed group differences in learning and recall. It was not possible to use ANCOVA to control for the effect of Age as relevant assumptions were not met; the allocation of participants to groups (Super-recogniser or Typical) was not random and the groups were significantly different on this variable. As an alternative way to investigate this, age is plotted against learning errors, 55min recall and 24hr recall in Figure 3.32. Since the aim is to investigate the association between the variables rather than look for group differences this analysis is performed for the aggregate group (N=134) made by combining the Super-recogniser and Typical groups. This ensures the analysis covers the widest possible range of each variable and maximises the sample size and hence statistical power.

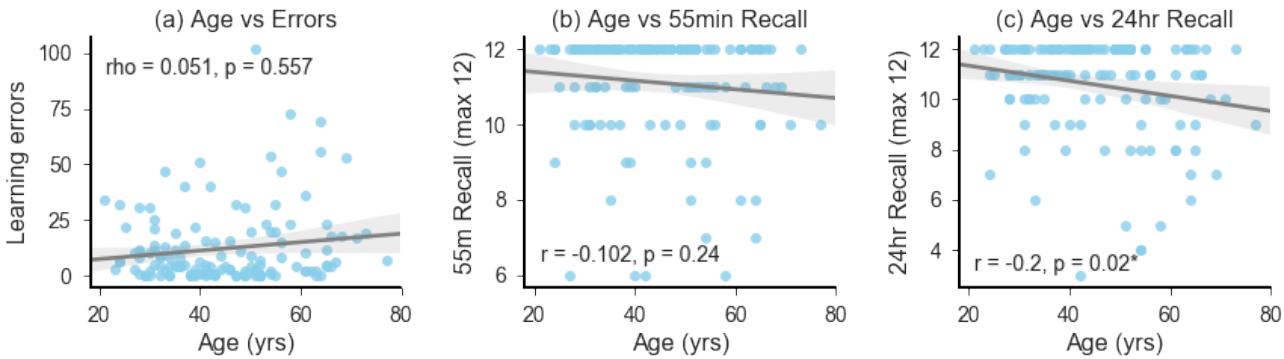


*Figure 3.32 Correlation of age with (a) learning errors, (b) 55min recall and (c) 24hr recall for the aggregate group containing all participants (N=134); shaded area is 95% confidence interval.*

The correlation with Age was small and non-significant for Learning errors (rho = .05, p = .56) and 55min recall (r = .10, p = .24, BF = 0.21), while for 24hr recall it was

also small but did reach significance before, but not after, Bonferroni adjustment (r = .20, p = .02, Bonferroni adjusted $p = .06$, BF = 1.54). These associations and the slope of the regression lines seem too small for the small age difference between the Super-recogniser and Typical groups (7.57yrs) to have driven the observed significant group differences in these three variables.

#### *3.5.3.5 Relationship between gender and VALMT variables*

In addition to the group differences in Age, there was also a significant difference in the gender split between the Super-recogniser and Typical groups. To investigate whether this gender imbalance was introduced by the VALMT exclusions or reflects a bias in the samples selected from the University of Greenwich face recognition database, the gender split was compared for all participants, before any VALMT exclusion criteria were applied (N = 495). This also showed a significant imbalance (Super-recogniser: 23M/100F, 81% female; Typical: 106M/211F/2O, 66% female; $X^2(2) = 10.09$, $p = .006$, BF = 0.03) which suggests that there are more females than males in the potential participants recruited and, more importantly, Super-recognisers are more likely to be female. Comparing the gender split before and after VALMT exclusions are applied showed that the bulk of the gender imbalance comes from the underlying sample recruited, rather than being introduced by the exclusions (before exclusions: Super-recogniser 81% female, Typical 66% female; after exclusions: Super-recogniser 89% female, Typical 68% female).

If super-recognisers are more likely to be female, then is the higher VALMT score of this group caused by being female, or being a super-recogniser, or a combination? One way to look at this is to check the impact of gender on VALMT scores within the Typical group. If females score higher on VALMT in this group, that would suggest having more females in the Super-recogniser group would raise that group's score even without the impact of any possible cross-domain cognitive advantage super-recognisers might possess. In fact, the gender differences for the Typical group were small and non-significant for all variables, although the difference for 55m recall did approach significance before Bonferroni adjustment: Learning Errors ($M_{Female}$ = 6.0, $M_{Male}$ = 10.0, $MWU = 606$, $p = .10$, $r = 0.22$, BF = 0.56); 55min Recall ($M_{Female}$ = 11.03, $M_{Male}$ = 10.30, $t(83) = 1.98$, $p = .05$, $d = 0.46$, BF = 1.29); 24hr Recall ($M_{Female}$ = 10.36, $M_{Male}$ = 9.78, $t(83) = 1.14$, $p = .26$, $d = 0.27$, BF = 0.42). The combination of p-values and small

effect sizes suggest that while an impact of gender on VALMT metrics cannot be ruled out, if one exists it can only account for a small part of the observed Super-recogniser vs Typical group differences.

#### *3.5.3.6 Distribution of learning errors across word-pairs and relationship to recall*

To investigate how delayed recall and errors made during learning vary across the 12 word-pairs, and to illustrate how these two variables relate to each other, Figure 3.33 shows the mean number of errors made for each individual word-pair plotted against the corresponding mean recall score at each delay, for the Super-recogniser and Typical groups.

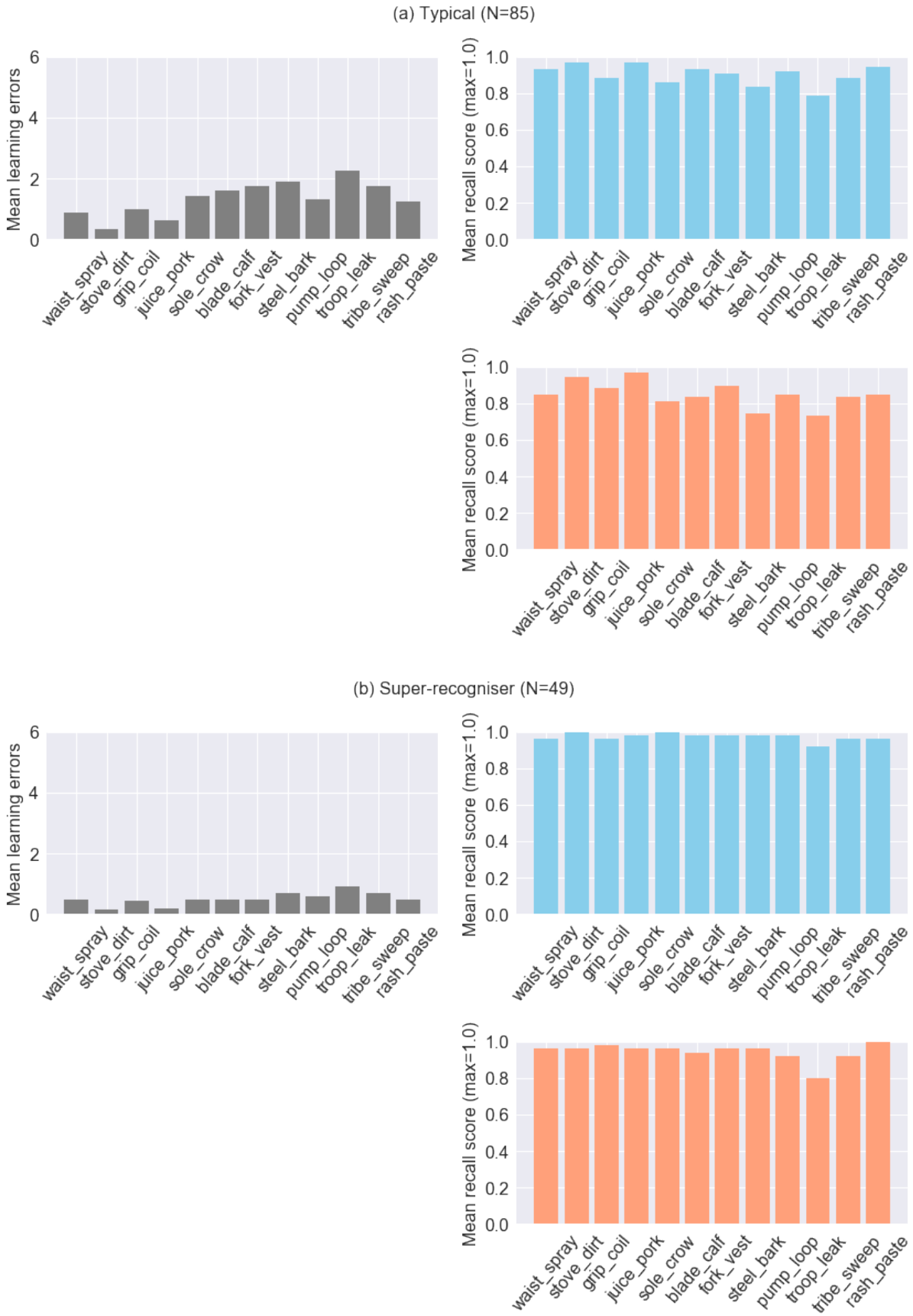


*Figure 3.33 Mean learning errors and delayed recall by individual word-pair for the (a) Typical and (b) Super-recogniser groups.*

Visual inspection shows that although the Typical group make more errors, the distribution of errors and recall across pairs is very similar to the Super-recogniser group. For both groups, those pairs which encounter fewest errors are recalled best while those pairs which resulted in most errors are recalled most poorly. There is variation in difficulty between pairs. This will be beneficial for distinguishing between performance levels, since some more difficult pairs are needed to challenge and differentiate between higher performers, and easier ones are needed to help differentiate between lower performers. No pairs encounter zero or a very high number of errors, or zero recall. This overall pattern matches that from previous experiments. However, unlike previous results, the Super-recogniser group in this experiment recall many pairs with, or close to, 100% success, indicating these pairs are too easy for these participants.

To investigate this association between errors and recall statistically at a per-pair level, the scatterplots in Figure 3.34 illustrate the relationship between the mean number of errors made learning each individual pair and the mean recall scores for that specific pair, for each group. These plots show one data-point for each of the 12 word-pairs.

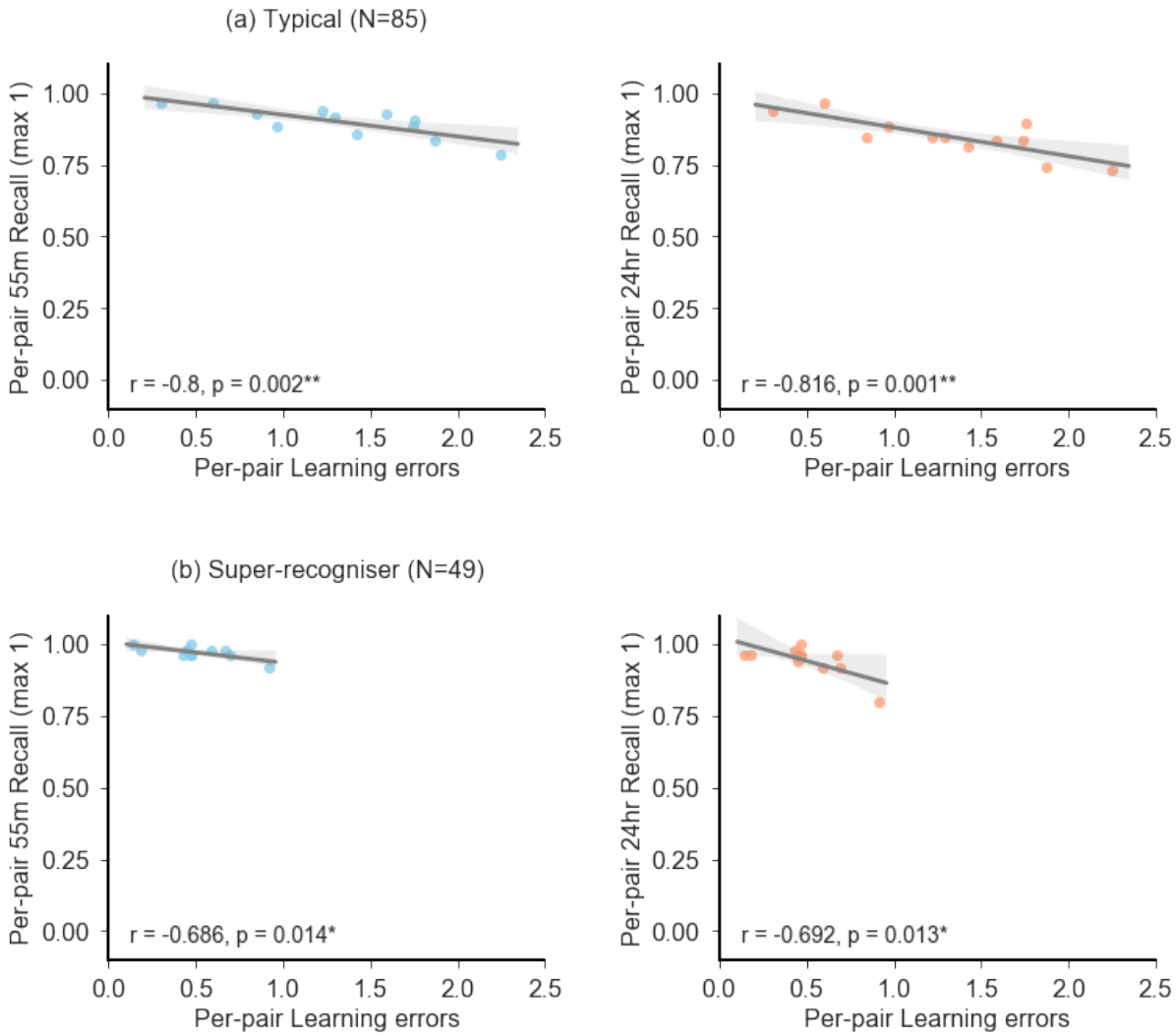


*Figure 3.34 Correlation of learning errors per word-pair with delayed recall per word-pair at 55m (left) and 24hr (right) for the (a) Typical and (b) Super-recogniser groups; shaded area is 95% confidence interval.*

There was a statistically significant and large or very large negative correlation between learning errors per individual word-pair and subsequent delayed recall per word-pair at both 55min and 24hr for the Typical group (55min r = -0.800, p = .002, BF = 27.0; 24hr r = -0.816, p = .001, BF = 37.0) and the Super-recogniser group (55min r = -0.686, p = .014, BF = 5.38; 24hr r = -0.692, p = .013, BF = 5.74). The effect size for the Super-recogniser group was smaller than for the Typical group, which is likely to be due to the large number of pairs which are recalled at 100% success rate by Super-recognisers, capping the variation in recall scores and distorting the relationship which the correlation measures. This is a form of ceiling effect. Overall, this analysis confirms the visual pattern seen in Figure 3.32, in which those pairs which encounter most errors are recalled most poorly.

### 3.5.3.7 *Ceiling effects*

Greenwich University advised that the participants in their database tend to be highly motivated, and to perform above representative population averages on most tests. It was noted that the Super-recogniser group in particular scored highly on the VALMT, and some pairs were recalled correctly by all participants in this group. This means it is likely that ceiling effects are impacting the results of some group comparisons and other statistical tests. To investigate this, Figure 3.35 shows the distribution of recall scores at 55m and 24hr for the Super-recogniser and Typical groups.

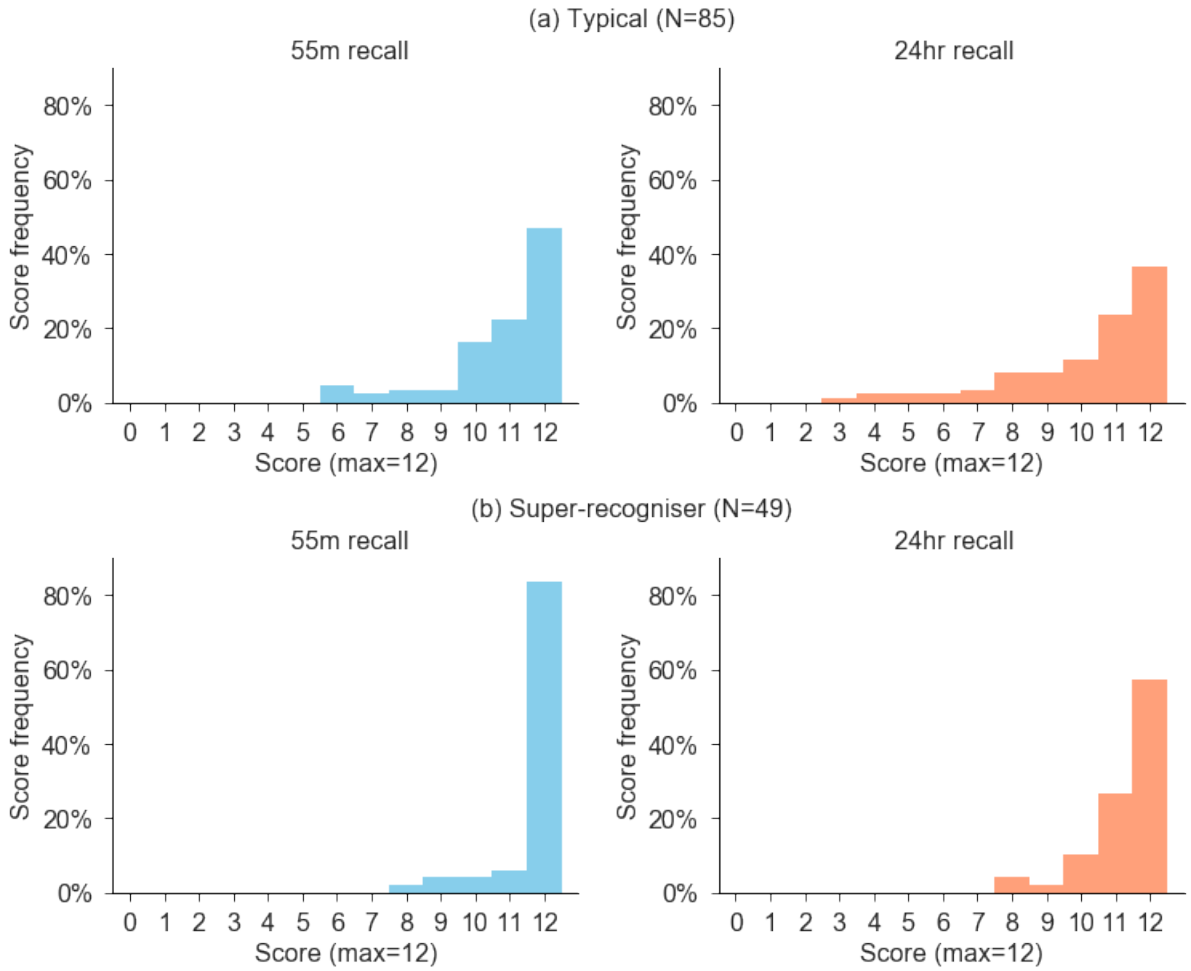


*Figure 3.35 Distribution of recall scores at 55min and 24hr as a function of group*

The distributions in Figure 3.24 show many participants scoring 'at ceiling' (achieving the maximum score), especially in the Super-recogniser group, confirming that ceiling effects will be present. In this situation, the scores of the highest scoring participants are effectively capped at the maximum score, while the scores of lower

performing participants are less impacted. This can lead to differential suppression of the mean score for the highest performing group, and a reduction in the observed difference in group means, leading to an under-estimation of the true effect size.

To further investigate the prevalence of ceiling effects in online VALMT studies Table 3.12 shows the number of participants achieving the maximum score for both the Typical and Super-recogniser groups in the current experiment, along with equivalent numbers for the Younger (18-30yrs) and Older (60yrs+) groups from Experiment 4, which reflect high and low performing groups recruited from the general population.

*Table 3.12 Percentage of participants achieving maximum score for delayed recall at 55m and 24hr as a function of experiment and group.*

| Factor | Expt 7 Super-recogniser | Expt 7 Typical | Expt 4 18-30yrs | Expt 4 60+yrs |
|---|---|---|---|---|
| **Younger** | | | | |
| N | 49 | 85 | 20 | 32 |
| Percent at max at 55m recall | 83.7% | 47.0% | 55.0% | 21.9% |
| Percent at max at 24hr recall | 57.1% | 36.5% | 50% | 15.6% |

The Super-recogniser group have by far the greatest proportion achieving maximum scores, so will have encountered the largest ceiling effects. However, apart from the Older (60yrs+) group from Experiment 4, all the other groups show proportions at ceiling which are a cause for concern and may have led to an underestimation of group differences.

### 3.5.4 Discussion

This experiment aimed is to investigate whether differences in memory performance detected by VALMT are specific to the verbal domain or correlate with a non-verbal form of memory, specifically with face recognition memory. Participants with high face recognition ability were recruited for a Super-recogniser group while participants with ability close to the population mean were recruited for a Typical group. These groups were then compared on VALMT performance and a clear group difference was identified, with the Super-recogniser group making fewer errors during learning, and scoring higher on delayed recall at both 55min and 24hrs.

Group comparisons for a range of demographics variables ruled out education and language as possible causes of performance differences, and also found that both groups

reported a very similar level of memory complaints. The mean age of the groups was different, with the Super-recognisers being younger. However, the relatively small age difference and the small effect of age on learning and recall meant this factor could also be discounted as a primary cause of the observed group differences.

The two groups were also different in gender makeup, with the Super-recognisers containing a higher proportion of females. This may reflect the general, although small, tendency for females to outperform males in face recognition tasks such as the CFMT (e.g. Bokak et al, 2016, $\eta_p^2 = 0.02$; meta-anlaysis by Herlitz & Loven, 2013, Hedges g = 0.36). There was evidence that females score higher than males on the VALMT generally, raising the possibility that the gender balance difference between the Super-recogniser and Typical groups may account for the differences in group VALMT performance. However, the size of VALMT performance difference between females and males generally was small. In a meta-analysis Hirnstein et al. (2022) found that the effect of gender on verbal recall was small for studies using the California Verbal Learning Test (CVLT; d = 0.42) and the Rey Auditory Verbal Learning Test (RAVLT; d = 0.39). These effects sizes are similar to VALMT results in the current experiment (55min recall, d = 0.46; 24hr recall, d = 0.27). The VALMT results therefore fit the general pattern in the literature. However, the small size of the effect combined with the size of the gender imbalance between groups meant this factor too, was unable to account for more than a small part of the observed Super-recogniser vs Typical group differences.

Overall, the results indicate that Super-recognisers do score higher on VALMT, even after allowing for possible demographic confounds. The fact that they score higher on both face-recognition tests and VALMT suggests that a common underlying cognitive ability, perhaps a common memory capacity, may partly drive results in both domains.

The relationship between errors during learning and subsequent delayed recall was the same as previous studies; those pairs which generate the most errors are recalled most poorly. This provides further evidence for the model proposed in the discussion of Experiment 4, in which the primary factor driving delayed recall is the number of errors made during learning, rather than underlying forgetting rate.

A significant feature of this experiment was the high level of recall, especially by the Super-recogniser group. Further investigation identified that many participants achieved the maximum score at recall, which will led to ceiling effects, with the group means for high performing groups being artificially suppressed. Effectively, the test is too easy for

the highest performers in this sample. Ceiling effects can be expected to suppress the higher group's recall score more than the lower performing groups' and thereby reduce observed group differences. Therefore, since a statistically significant difference was found between Super-recogniser and Typical groups on delayed recall at both delays with ceiling effects present, any changes to reduce ceiling effects would be expected to make these differences larger rather than smaller, so the general conclusion that super-recognisers perform better on VALMT is still valid.

The VALMT is targeted at identifying subtle memory deficits in the general population caused by conditions such as early stage AD or epilepsy. It is not intended to be used to differentiate high performing individuals with no memory deficits. The Super-recogniser group are, by selection, an above average performing group. Therefore, for general use of the VALMT in the way it was intended it is perhaps not an issue that a significant number of Super-recognisers score at ceiling. However, a comparison across experiments showed that the proportion of participants scoring at ceiling in the higher performing group in previous experiments (the 18-30yr Younger group), while not as high as the Super-recognisers in the current experiment, was still a cause for concern. The impact of ceiling effects on results and possible changes to the VALMT to reduce ceiling effects will be discussed in the next chapter (Chapter 4).

### 3.5.5 Conclusion

This experiment has shown that the differences in memory performance detected by VALMT are not unique to the verbal domain; they also correlate with face recognition memory, a non-verbal domain. This suggests that VALMT scores are partly driven by a common underlying cognitive ability, perhaps a common memory capacity.

Testing with high performing groups in this experiment highlighted that ceiling effects are present in VALMT results for higher performers. Further work will be required to quantify and evaluate the effect of these, and to investigate modifications to VALMT to reduce these.

## 3.6 General discussion

This chapter reports four experiments that evaluate a new online version of VALMT and use it to address several questions and issues raised by the face-to-face experiments presented in Chapter 2.

Across all four experiments the new online system worked as expected, with no technical problems or procedural issues reported, despite a wide range of devices being used (e.g. phones, laptops, tablets). This shows that an online version can work well for all ages in the range from 16 to at least 82, and will make it possible to run experiments that recruit without geographic restrictions and test at multiple delays without the need to travel for face-to-face testing, and also allowing large numbers of participants to be tested without the need for a large number of trained researchers.

However, although online operation makes it easier to recruit participants by widening the potential pool and eliminating the need for travel, it did lead to significant attrition, with some participants dropping out without completing all stages, particularly where participants were recruited from the general population. While analysis showed the dropouts were unlikely to have impacted any of the group differences measured in this series of experiments, it would be desirable to reduce this in future. It should be noted that no financial reward of any sort was offered for any of these experiments. However, for Experiment 3 the participants who completed all stages did receive participation credits for Goldsmiths College's research participation scheme, which is an essential requirement for their undergraduate degree. The resulting dropout rate was lower than for the remaining experiments, suggesting that offering some form of reward could help reduce attrition.

Participants in all four experiments were asked to complete the delayed tests as close as possible to specific delays (55min and 24hrs), but were also encouraged to complete the tests whenever possible if they were unable to return to the website at the correct time for any reason. This approach was used to maximise data capture and to allow the impact of varying the test delays to be investigated. However, the fact that so many participants completed one or more of their delayed tests outside the normal window is a potential issue for future use. Although the instructions asked participants to use a reminder of some sort (alarm on mobile phone, calendar or diary entries etc.) to ensure they completed the delayed tests at the correct time, building an automated reminder mechanism into the system would be beneficial. This could either send an email or an SMS message to remind participants to return to the website for their delayed tests. Such a reminder system should also help reduce attrition, by avoiding the situation were a participant completely forgets to complete the delayed tests. Developing the system to do this is made more complex by privacy issues since this will involve storing phone

numbers or email addresses which tie data records to identified individuals; this will be beyond the scope of this PhD.

Across the four experiments the results replicated the general pattern seen in the face-to-face experiments. Specifically, it was possible to identify differences in the memory performance of Younger and Older groups by 55 minutes, these group differences were driven largely by slower learners within the Older group who require more trials to reach criterion and display more rapid forgetting, and this Older_Slow group also reported more memory complaints. This provides a level of concurrent validity, suggesting that the face-to-face and online version are measuring the same thing, and also strengthens the evidence base for the important result that there is a subset of healthy older participants who learn more slowly and forget more rapidly. However, while the same general pattern was observed, the size of the group differences was lower than observed in the face-to-face experiments. Analysis showed that the online Younger groups performed very similarly to the equivalent groups in face-to-face experiments, and that the smaller group differences were largely driven by the Older groups performing better in the online experiments. Multiple possible causes were investigated, with evidence suggesting this is most likely due to a combination of two primary factors.

First, the online version imposes a 20min time limit for the learning phase, something which was not done for the face-to-face experiments. This excluded a small number of participants who were the poorest learners. Including these participants back into the analysis of learning performance increased the group differences between the Old_Slow group and the Younger and Older_Fast groups. Similarly, re-analysing the face-to-face Experiment 2 with a retrospective 20min time limit enforced reduced the group differences seen in that experiment for both learning performance and delayed recall. Together, this suggests that the learning stage of the VALMT is able to identify those with the greatest memory deficit without the need to complete any delayed recall testing. This result is important clinically, as it may allow for rapid diagnosis of a memory deficit while minimising both effort on the part of the clinician and stress on the part of the patient.

Second, the online version employs a much simplified procedure which eliminates the relatively complex interleaving of learning and testing that was used in the face-to-face version. It is possible that the interleaving provides the opportunity for interference to impact later delayed recall. The fact that the Younger group performed very similarly for both versions suggests they are not impacted by this. However the Older groups may

have been impacted, which would then account for part of the improved performance seen in these groups in the online version. Unfortunately, the better learning performance of the online Older group compared to the face-to-face equivalent group means this conclusion must remain provisional; it is also possible that the Older group in the online experiment were simply a higher performing group and this drove their improved scores rather than the removal of interference. Such a difference could be caused by random sampling effects, or could be due to a selection bias, with poorer performers more likely to drop out during the online test than the face-to-face test. Further data will be required to choose between these explanations for the lower performance of the Older group in the online version.

Across all four experiments the link observed in face-to-face experiments between learning errors and subsequent delayed recall was replicated. Those people who make the most errors recall the least material. Importantly, the additional statistics gathered by the online version allowed this relationship to be investigated at a more detailed level, looking at the errors and subsequent recall for each word-pair. This granular data showed that it was not the total number of errors made that was most important, but rather where those errors were made. The word-pairs which generate the most errors are the ones which are recalled most poorly. The fact that these per word-pair errors appear to be the strongest predictor of subsequent recall performance across all ages suggests the underlying mechanism is the same for all ages, rather than being something which arises only in the Older_Slow groups. Instead, it is the increased number of errors made during learning by these slower learners then drives this universal mechanism and leads to poorer recall.

Although the link between errors during learning and subsequent recall is strong, it also remains possible that the Older_Slow group also have a greater underlying forgetting rate, independent of the influence of errors, so that two mechanisms are involved in their poorer recall. While it was not possible to investigate this using the face-to-face version, the detailed data gathered for every response by the online version makes such analysis possible. Some word-pairs are learnt to criterion without any making any errors. Such ‘errorless’ pairs are all presented and recalled the same number of times, and generate no errors. This makes their learning very well matched across participants. The forgetting rates for these errorless pairs provide a good indicator of participants’ underlying forgetting rate while controlling for the confound of errors during learning. Analysis of these rates showed no evidence of greater underlying forgetting rates in the Older-Slow

group in comparison to the Older_Fast group. This provides strong evidence that the poorer delayed recall displayed by the Older_Slow groups is caused by a learning deficit, which causes them to make more errors, which in turn leads to poorer recall. It seems that, at least within the timeframe of 24hrs, their main issue is a learning deficit. It remains possible that testing recall at longer delays, such as 7 days as used by Weston et al. (2018), may identify a difference in underlying forgetting rates over such extended delays.

If the poor recall of some Older participants is down to the greater number of errors they make during learning, then obvious follow-up questions are why they make more errors and what the significance of this. While there are a number of possible explanations, the evidence from VALMT experiments suggests this is primarily down to a cognitive deficit which impacts learning and that this may indicate the earliest stages of AD or another form of dementia.

The data from Experiments 4 and 5 allowed memory performance across the lifespan to be investigated, covering the age range from 16 up to 82 years. Across learning, recall, memory complaints and forgetting rate it was clear that memory performance does decline in later years. The size of the dataset means the analysis is relatively coarse, so the age of onset of decline cannot be quantified precisely. However, it seems that memory performance starts to decline by the 57-69yr age band. The only other studies identified in the literature which investigated memory performance over such extended delays (up to at least 24hours) and for such a wide range of ages were Davis et al. (2003) and Huppert and Kopelman (1989). Davis et al. found evidence of accelerated forgetting after 24hr in their oldest age band (76-90yrs). Huppert and Kopelman, after careful matching of performance at 10min, found evidence of accelerated forgetting between 10min and 24hr in their 38-64yr middle-aged group. The current study therefore suggests that a decline in performance may start earlier than found by Davis et al., and at an age more in line with Huppert and Kopelman. However, a larger sample size will be required to validate this observation. At the other end of the age range, there was no evidence of any lower performance in the 16-17yr age group, compared to the 18-30yr age group, so it seems that memory of the type measured by VALMT has already approached its peak level by this age, a result which is again consistent with Huppert and Kopelman.

The experiments in this chapter are the first VALMT experiments to test recall at the longer delay of 24 hours. This was intended to look for any late onset forgetting that might not be apparent when testing at 55mins. The results showed relatively little

forgetting between 55min and 24hr, and there was no significant late onset forgetting identified in younger or older groups. This was likely influenced by the use of a repeated recall, with the same pairs being recalled at both intervals. It may be that the delay for the second delayed test will need to be extended to several days in order to detect late-onset forgetting. However, it could also be that the tougher learning to criterion procedure of the VALMT may cause any recall impairment to manifest sooner than it does with other tests with less challenging learning procedures, which then means those who would have shown late-onset forgetting in other published studies can be identified earlier using the VALMT. Answering this will require further testing with older participants at longer delays.

The VALMT measures learning and memory performance in the verbal domain, using pairs of unrelated words as stimuli. While the underlying ability driving the test results could be specific to the verbal domain, it could also be a more generic ability which spans domains. By comparing the performance of participants on face-recognition memory and VALMT, Experiment 6 suggests that performance may be partly driven by a shared underlying cognitive ability, perhaps a cross-domain memory capacity. This would fit with many modular theories of memory function which postulate a common indexing function in the hippocampus, combined with domain specific processing and storage areas in the wider medial temporal lobe and neo-cortex (e.g. Moscovitch, 2008; Alvarez & Squire, 1994; Nadel & Moscovitch, 1997). In these models the hippocampus would be important across domains, leading to the cross-domain correlations in memory performance. In contrast, the domain specific regions and communication between these and the hippocampus would drive the cross-domain differences. However, there are other explanations and possible confounds. For example, the difference in task type (cued-recall for VALMT; recognition for face memory) could have influenced results. Further research with carefully designed comparisons will be necessary to build on and elucidate this initial result.

The highest performing group across all four experiments was the Super-recogniser group from Experiment 6. This group performed very highly on face-recognition memory and also very highly on VALMT. They all obtained the maximum score on the GFMT, and many also scored at ceiling on the VALMT delayed recall tests. Such high scoring can lead to ceiling effects, in which the true ability of high performers is under-estimated, and the mean performance for such groups are artificially suppressed leading to reduced group differences when compared with lower performing groups. While this is most

noticeable for the Super-recogniser group, analysis of other VALMT experiments showed that the proportion of participants achieving the maximum score is a cause for concern in higher performing groups in general; for most experiments this means the 18-30yrs Younger group. It will be difficult to completely eliminate both ceiling and floor effects when measuring performance for a mixture of poor and high performing individuals at both short and long delays; preventing the highest performers scoring at ceiling at the short delay while also preventing the lowest performers scoring at floor at the longest delay is unlikely to be possible. This raises the question of what size of ceiling and floor effects are tolerable. As no concrete guidelines could be found in the existing literature, this question will be examined empirically in detail in the next Chapter.

# 4 Chapter 4 – Avoiding and mitigating ceiling effects in ALF research

## 4.1 Introduction

The analysis in Chapter 3 highlighted that ceiling effects may be influencing the results of VALMT studies. This reflects a concern in accelerated forgetting research more widely that methodological flaws, and ceiling effects in particular, may lead to a failure to detect differences between groups at short delays such as 30 minutes (Elliott et al., 2014; Cassel et al., 2016). ALF studies often involve testing at 30 minutes and at least one extended delay. The extended delay can be as long as several weeks. To avoid all participants scoring at, or close to, zero (*scoring at floor*) at such extended delays it is common to ensure material is initially learnt to a high criterion. While this helps to avoid participants scoring at floor at long delays it also leads to high performance at shorter delays, with the potential for ceiling effects to distort results.

Ceiling effects are distortions to performance measurements that can occur when there is a maximum score associated with a task. This maximum score enforces an upper limit, or *ceiling*, on the score that any individual participant can achieve. When someone does score at ceiling the task is too easy for them, and their true ability is not being measured, reflecting a loss of information. If multiple participants in an experimental group score at ceiling then the mean score for that group will be lower than it would have been if there was no maximum. The range of scores within the group will also be reduced, along with other measures of variation such as standard deviation.

Ceiling effects can distort the results of statistical comparisons between groups. Depending on the nature of the comparison, this can lead to the failure to detect a difference where one exists (*false negative*), or the detection of a difference where none exists (*false positive*). For example, if there are two groups in a study, any ceiling effect will impact the higher performing group more than the lower performing group. The mean score for the higher performing group at a short delay will be reduced more than that of the lower performing group, reducing the difference in means between the groups. In this case, a comparison of the mean scores at a short delay may fail to get a statistically significant result, even though there is a genuine underlying difference between the ability of the groups (false negative). In contrast, when forgetting rates are being compared, the reduction of the higher performing groups scores at the short delay can reduce the apparent forgetting rate after that delay. This can cause the forgetting rate to

look lower than that for the lower performing group, leading to a false claim of accelerated forgetting in the lower performing group (false positive). Similar arguments apply to floor effects, typically occurring at longer delays, when many in the lower performing group score at floor.

The impacts of ceiling and floor effects are seldom explicitly discussed in statistics text books aimed at the social sciences, and when they are, the discussion is generally limited to a qualitative explanation such as that provided above together with a simple statement that ceiling and floor effects should be avoided. At the time of writing it has not been possible to identify quantitative analysis or recommendations in any text book of this sort.

Unlike statistical text books aimed at the social sciences, the wider statistics literature does discuss ceiling and floor effects and techniques to deal with them, although the terminology used is both different and more explicit. The key term used is *censoring*. A sample is said to be *censored* where a maximum or minimum score is enforced resulting in a loss of information; any score that would have been above or below these limits is set to the limit value. The limit score is known as the *censoring point* or *truncation threshold*. The *impact of censoring* is then equivalent to the ceiling effect and floor effect terminology more common in the social sciences. These terms are used interchangeably in this chapter.

The nature of experimental designs in accelerated forgetting research, with testing performed at both short and long delays, makes it extremely difficult to completely avoid both ceiling effects at short delays for high performing groups and floor effects at long delays for low performing groups. This makes it important to understand how much ceiling and floor effects distort results, and what level of these effects can be tolerated in practice, rather than trying to completely eliminate them. None of the literature reviewed provided any such guidance. To overcome this gap in the literature, the aims of the analysis in this chapter are to develop design guidelines that ensure censoring does not invalidate results in the specific experimental design which is typical for ALF research, and to investigate techniques to better analyse data where censoring has occurred.

In a real world experimental setting it is not possible to definitively know the true population means, as that would require testing the entire population(s) in question. This means it is very difficult to accurately assess how much influences such as ceiling and floor effects have distorted results. It is also not practical to run large numbers of experiments with different populations to understand how factors like the true effect size

or sample sizes mediate these effects. For this reason this chapter will use a computer simulation approach, in which statistical software is used to draw samples from simulated populations defined with specific means and variances. This allows the impact of ceiling and floor effects to be accurately assessed for a wide range of sample and effect sizes.

## 4.2 Impact of censoring on a single group

The impact of censoring is first illustrated by looking at the simplest case, a single group with a single experimental condition. This analysis assumes that for a measure that has no minimum or maximum scores then the distribution of scores can be expected to be normal. Figure 4.1 illustrates what such a distribution of scores looks like, first with the censoring point (maximum score) set so high that it has no effect (Fig 4.1(a)), and then with the censoring point reduced in increments.

For this example, 500 random values were drawn from a population with mean 10.0 and standard deviation 1.0. These specific mean and SD values were chosen as they reflect typical values in published ALF studies (e.g. Butler et al., 2007; McGibbon & Jansari, 2013), and these values are also used as the default values for all following analyses. In most real-world experiments the actual sample size will be much smaller than 500; this number was chosen to reduce the impact of sampling error, resulting in a smoother distribution, and making it easier to focus on the impact of censoring. The red vertical line in the figure identifies the maximum score possible i.e. the censoring point. As the censoring point is reduced from 14 to 7 the impact of censoring becomes greater. For this simple model, any score which was above the censoring point in the original distribution is set to the censoring point value. For example, with the censoring point set to 11 the model changes any scores above 11 to 11.

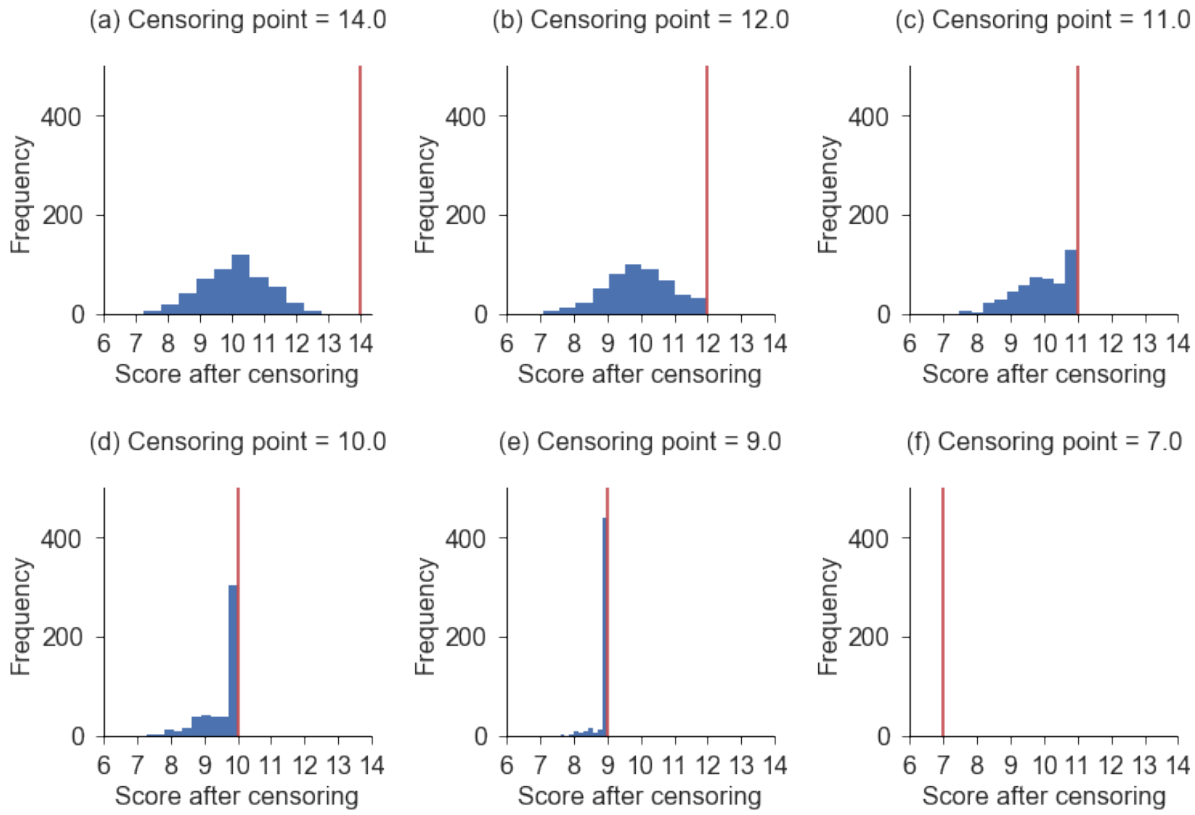


*Figure 4.1 Impact of censoring on the distribution of scores for a single group. Population mean 10.0, SD 1.0, sample size 500. Red line indicates censoring point.*

The distribution is initially normal, centred on the population mean score of 10.0 (Fig 4.1(a)). As the censoring point is reduced the distribution becomes progressively less normal. When the censoring point reaches 7 all participants are scoring at maximum, so the distribution is a single vertical bar of height 500 (Fig 4.1(f)). The mean score for the group reduces as the censoring point is lowered, changing from the initial value of 10.0 to a final value of 7.0. The total variation in the distribution also reduces; the distribution is progressively 'squashed' into a smaller range of scores and any measure of variation such as standard deviation will also be reduced.

## 4.3 Impact of censoring on the difference between two group means.

Next the model was expanded to draw samples from two different populations, to simulate two experimental groups. The impact of censoring is illustrated in Figure 4.2. In this analysis, samples were drawn from populations with means of 11.0 (blue bars in the figure) and 10.0 (green bars). The standard deviation for both populations was 1.0. These values equate to a Cohen's d effect size of 1.0, which is 'large' by convention, and is a difference that any well designed experiment should to be able to detect.

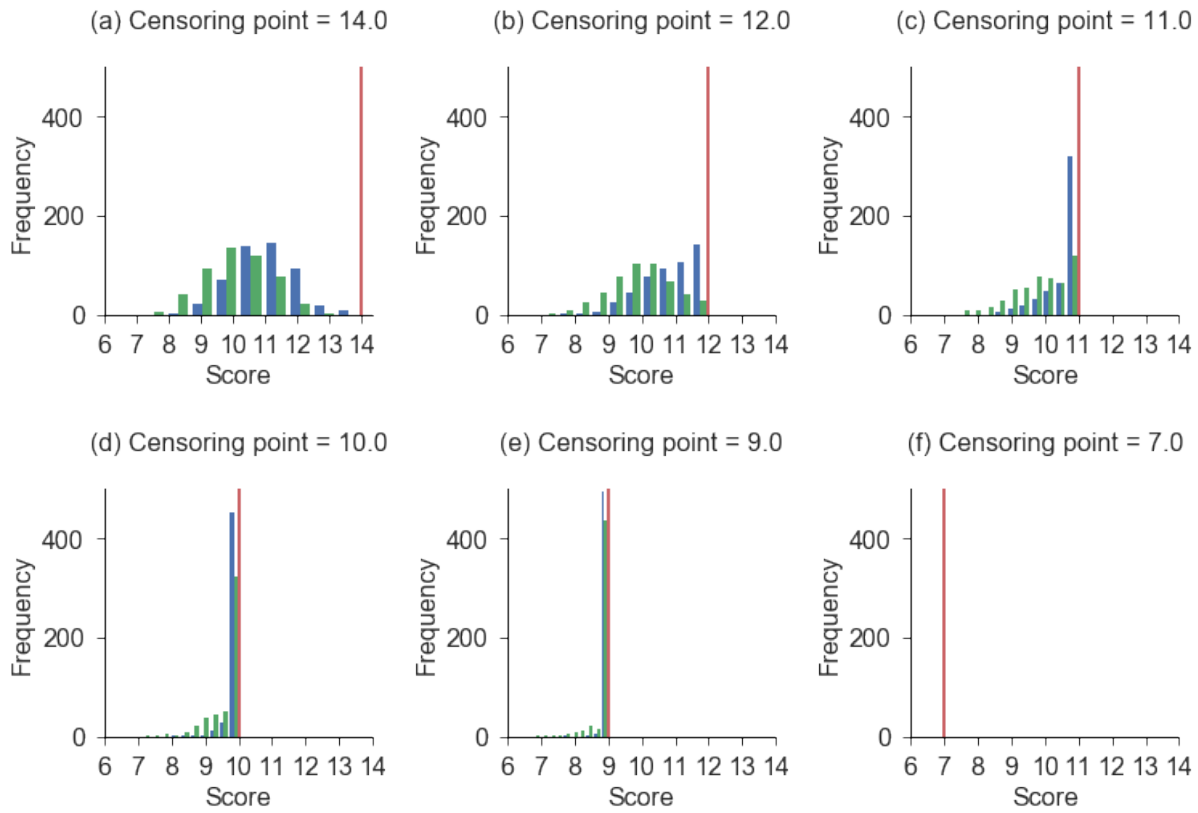


*Figure 4.2 Impact of censoring on the distributions of scores for two groups. Population means 11.0 & 10.0, SDs 1.0, sample size 500. Red line indicates censoring point.*

The censoring point was gradually reduced from 14 to 7, as before. Visual inspection of Figure 4.2 shows that while the impact on each group's distribution is eventually the same, censoring impacts the distribution with the higher mean first. This differential impact to the distributions is visually most noticeable with the censoring point set to 12.0 (Fig 4.2(b)), where the distribution for the group with the lower mean (green bars) is still approximately normal, while the distribution for the higher performing group (blue bars) is now heavily skewed.

Although Figure 4.2 provides some intuitive understanding of the problem, it does not provide quantitative data about the effect of censoring on the difference between group means. The next analysis obtained such data by calculating the difference between the sample means before and after censoring, and then running the same simulated experiment 10,000 times to average out the effects of sampling variation. The sample size was reduced to 25 for each group, which is more representative of a real-world ALF experiment. The following sequence of steps was performed:

1. Draw 25 samples each from two populations, with means 11.0 and 10.0, SD =1.0
2. Calculate the difference between the means of the two sample groups (*mdiff_uncensored*)
3. Censor the samples using a censoring point of 14.0, and again calculate the difference between the means of the two sample groups (*mdiff_censored*)
4. Record the change in the difference between group means from the uncensored to the censored condition (*mdiff_uncensored – mdiff_censored*). This is the impact of censoring for a single simulated experiment; the amount by which the observed difference in group means changes due to censoring at a particular point.
5. Repeat steps 3 and 4 for multiple censoring points, between 14.0 and 7.0. This provides the impact of varying the censoring point within a single simulated experiment.
6. To provide an unbiased estimate of the impact of censoring, free from the sampling error present in any single experiment, repeat steps one to five 10,000 times, to simulate running 10,000 experiments, and providing 10,000 estimates for the impact of censoring for each censoring point value between 14.0 and 7.0.
7. For each censoring point value calculate the mean of the 10,000 recorded impact estimates.
8. Plot mean censoring impact against censoring point value.

Figure 4.3 shows the resulting plot of censoring point against the impact of censoring on the observed difference between group means. A censoring point of 14 has no impact on the difference between the groups means (impact = 0%); this is because no individual scores are high enough to be censored. At the other extreme, a censoring point of 8.0 reduces the difference between group means by 100%, eliminating all difference. This is because every participant, irrespective of group, scores at ceiling, so the group means are equal.

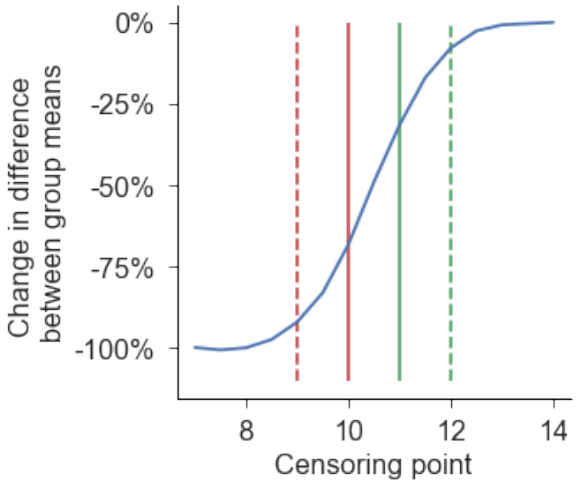


*Figure 4.3 Impact of censoring on the difference between group means. Population means 11.0 & 10.0, SDs 1.0, sample size 25, 10,000 simulations. Solid lines indicate population means (11.0, green; 10.0, red); dashed lines show 1SD above higher and below lower population means respectively.*

The solid vertical lines on this plot indicate the population means (11.0, green; 10.0, red). The dashed lines represent 1 SD above the higher mean and 1 SD below the lower mean; the SD for both populations is 1.0, so these bars are at 11 + 1 = 12 and 10 – 1 = 9. The plot shows that when the censoring point is greater than 1SD above the higher population mean, censoring has little impact (change in difference between group means is close to 0%). This suggests that a reasonable rule of thumb to avoid a significant ceiling effect is to ensure any censoring point is above this threshold. This makes sense as very few participants will have scores higher than 1SD above the higher population mean (approx. 16% of the higher mean group and only 2.3% of the lower mean group), so the censoring does not influence many scores.

As the censoring point reduces from 12 to 9 the impact on the difference between group means increases almost linearly. At a value below 9 any further reduction has little additional impact. This again makes sense as very few participants will have scores lower than 1SD below the lower population mean, so any further reduction in the censoring point does not influence many additional scores.

Two key conclusions from this analysis are:

- to prevent significant distortion of the observed difference between group means, keep the censoring point higher than 1SD above the population mean of the higher performing group; this equates to a maximum of 16% of the higher performing group scoring at ceiling.

- if the censoring point is in the range between 1SD above the higher mean and 1SD below the lower mean then it is worth trying to increase the censoring point; any change made will have a nearly linear impact on the observed difference between group means

## 4.4 Impact of censoring on the p-value for a t-test applied to independent group means

The previous section (4.3) investigated the impact of censoring on the observed difference between group means. However, section 4.2 highlighted that censoring will also reduce the variance in sample distributions. Since many statistical tests such as the popular t-test are dependent on both the between-group variance and the within-group variance the impact of censoring on any test of statistical significance will be more complex.

The visual analysis of section 4.2 also shows that censoring changes the shape of the sample distributions, making them non-normal. This means any test that has normality of distributions as a requirement should not be used. The t-test does have an assumption of normality, but is also relatively robust to non-normality. This means it can often be used successfully with non-normal distributions. Analysis of many published studies shows that researchers often continue to use a t-test even in the presence of censoring, and therefore non-normality. This may be based on the robustness of the t-test to non-normality, or it may be due to oversight, but whatever the reason it will be beneficial to understand the impact of censoring on t-test results.

This section expands the simulation procedure of the previous section (4.3) such that in addition to the difference in group means, the p-value of the t-test was also recorded for each simulation at each censoring point. The same populations, censoring points and sample sizes were used (population means = 11.0 and 10.0, SDs = 1.0, sample sizes = 25, censoring point varied from 14.0 to 7.0, results averaged across 10,000 simulations for each set of parameters). The mean change in p-value caused by censoring is plotted against censoring point value in Figure 4.4. Before censoring, the mean t-test p-value is .014. There is a genuine difference of 1.0 between the population means, and this mean t-test p-value indicates that an experiment would, on average, detect this difference when no censoring is present. Figure 4.4 shows how censoring changes this p-value.

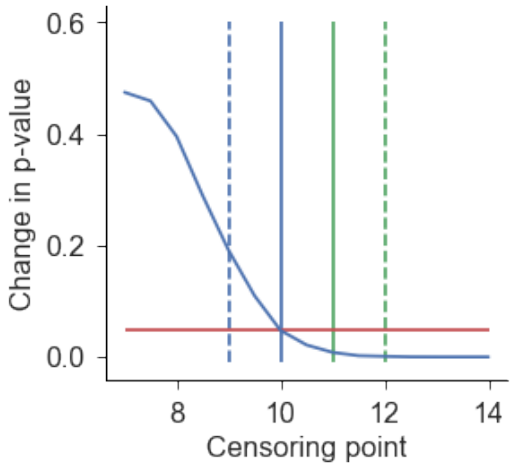


*Figure 4.4 Impact of censoring on independent means t-test p-value. Population means 11.0 & 10.0, SDs 1.0, sample size 25, 10,000 simulations. Solid lines indicate population means (11.0, green; 10.0, blue); dashed lines show 1SD above higher and below lower population means respectively. Solid red line indicates an increase in p-value of .05.*

Any censoring point above 12.0 has no impact on the mean p-value. As the censoring point is reduced below 12.0 the mean p-value starts to increase. At the other extreme, a censoring point of 7.0 adds .5 to the mean p-value. Importantly, a censoring point of 10.0 adds approximately .05 to our mean p-value. This means that a censoring point equal to the mean of our lower performing group population will lead to our t-test, on average, failing to find a significant difference between our groups due to censoring alone, assuming the standard p = .05 criterion for statistical significance (alpha).

The impact on the mean t-test p-value starts increasing rapidly when the censoring point is below 11.0, the population mean of the higher performing group. Comparing this with the results from the previous section (4.3), as the censoring point is lowered it starts to impact the difference between group means sooner than it impacts the mean t-test p-value. Specifically, censoring starts to impact the difference between group means when the censoring point is approximately 1SD above the higher population mean, and does not impact t-test p-values until it is equal to the higher population mean. This may be due to the interaction of the impact of censoring on the between group and within group variances. The impact of censoring on within group variance appears to reduce the impact on t-test p-value.

If the difference between groups was greater the t-test p-values would be lower, and perhaps the impact of censoring on p-values would be less significant. To investigate whether the effect of censoring is dependent on the size of group differences in this way

Figure 4.5 illustrates the effect of varying the difference between the population means. Here all parameters are the same as the previous analysis, except that population means of 11.5 & 9.5, 11.0 & 10.0, and 10.75 & 10.25 are used to generate population mean differences of 2, 1 and 0.5 respectively, equating to Cohen's d effect sizes of 2.0, 1.0 and 0.5.

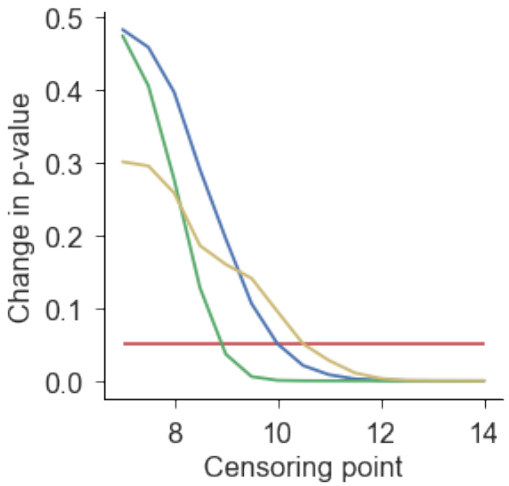


*Figure 4.5 Impact of censoring on independent means t-test p-value for multiple population means 11.5 & 9.5 (green), 11.0 & 10.0 (blue), 10.75 & 10.25 (yellow), SDs 1.0, sample size 25, 10,000 simulations. Solid red horizontal line indicates an increase in p-value of .05.*

The blue curve is a repeat of the previous analysis. The yellow curve illustrates the impact of censoring when the difference between population means is smaller (effect size 0.5, 'medium' by convention), and the green curve illustrates the impact with a higher difference in population means (effect size 2.0). As predicted, with a smaller difference between the population means (yellow) the impact of censoring on mean p-values starts at a higher censoring point, at around 1SD above the population mean of the higher performing group (11.75). By contrast, with a larger difference between populations the mean p-value is not impacted until the ceiling has been reduced to around 9.5, which is the mean of the lower performing group. In summary, the impact of censoring on mean p-values is more pronounced when the difference between populations is smaller. Furthermore, with a smaller difference between populations the t-test p-values will already be higher before any censoring, which makes the test of statistical significance more vulnerable to the impact of censoring, further exacerbating the impact of censoring on the ability to detect a statistically significant difference using a t-test.

A further factor influencing the p-value obtained from a t-test is sample size. In general, larger sample sizes make the statistical test more powerful. How does sample

size interact with censoring? It might be predicted that studies with larger sample sizes would be less vulnerable to censoring. To investigate this, the analysis was repeated for sample sizes of 10, 25, 50 and 100, with the difference between population means set to 1.0, and all other parameters the same as in the previous analyses. The results are illustrated in Figure 4.6.

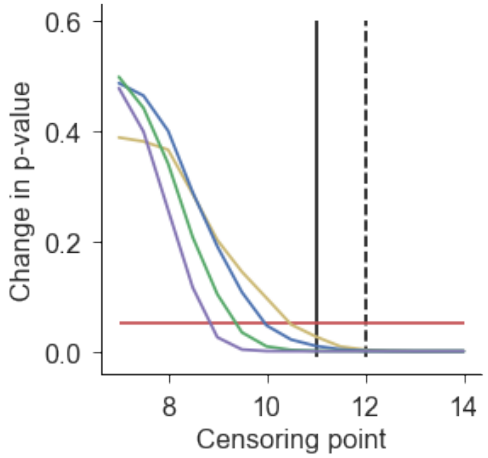


*Figure 4.6 Impact of censoring on independent means t-test p-value for sample sizes of 10 (yellow), 25 (blue), 50 (green) and 100 (magenta). Population means 11.0 & 10.0, SDs 1.0, 10,000 simulations. Vertical lines show the higher population mean (solid) and 1SD above higher population mean (dashed) respectively. Solid red horizontal line indicates an increase in p-value of 0.05.*

As predicted, as sample size increases the t-test becomes less vulnerable to censoring, with the impact on mean t-test p-value starting at a lower censoring point. With a sample size of 10 the impact starts with the censoring point at around 11.5, while for a sample size of 100 it starts with the censoring point at around 9.5. To prevent censoring adding to t-test p-values the censoring point should be kept above 11 with sample size 25, and 11.5 with sample size 10. These equate to 50% and 31% of the high performing group scoring at ceiling, respectively.

Small sample sizes are more typical of neuropsychological studies. For a sample size of 10, when the censoring point is set at around 10.5, which is midway between the population means, censoring adds .05 to the mean t-test p-value, and thus makes a t-test non-significant on average due to censoring alone. At this value, 50% of all participants would score at ceiling: 69% of the higher performing group and 31% of the lower performing group. With a sample size of 25, the censoring point must be lowered a little further, to 10.0, to add .05 to the mean p-value. At this value, 50% of the lower performing group and 84% of the higher performing group would score at ceiling, or an overall figure of 67% of all participants.

The key conclusions from this analysis are:

- On average, the difference between group means is more sensitive than t-test p-values to the impact of censoring, with the impact starting when the censoring point is at a higher value. This means following the rule of thumb for censoring threshold identified in section 4.3 (i.e. to prevent significant distortion of the observed difference between group means keep the censoring point at least 1SD above the population mean of the higher performing group; 16% of the higher performing group score at ceiling) will also, on average, avoid censoring impact on t-test p-values.
- However, if some distortion of the estimate of difference in population means can be tolerated then a less stringent guideline can be used while still preventing the impact of censoring on t-test p-values. With sample and effect-sizes typical of ALF studies (10-25 and 1.0 respectively), an appropriate target is to keep any censoring point high enough such that no more than 31% (sample size = 10) or 50% (sample size = 25) of the higher performing group are at ceiling. This provides an alternative rule of thumb for study design.
- With sample and effect sizes typical of ALF studies (10-25 and 1.0 respectively), if 69-84% of the higher performing group score at ceiling then censoring alone will add .05 to p-values on average and prevent detection of a large effect (Cohen's d = 1.0)

## 4.5 Impact of censoring on the likelihood of a study detecting an effect (power)

The previous section (4.4) investigated the impact of censoring on t-test p-values. However, when designing a study a more conventional criterion is statistical power, which is the likelihood of detecting an effect when one is present. It is typical to target a minimum power of 80%. If power is much lower than this then the researcher risks investing a large amount of energy and resources on a study that has little chance of detecting an effect, even when one is present.

To investigate the impact of censoring on power the process from the previous analysis was extended to include recording the result of the t-test significance check (significant or non-significant) for each censoring point, using the standard p = .05 as the significance threshold. This was repeated using population means with differences of 2.0, 1.0 and 0.5

(means 11.5 & 9.5, 11.0 & 10.0, and 10.75 & 10.25). Other values were as used previously (SDs = 1.0, 10000 simulations, sample size 25).

The results are illustrated in figure 4.7, with the percentage of simulations that result in a significant result (equivalent to statistical power) plotted against censoring point for each difference in population means.

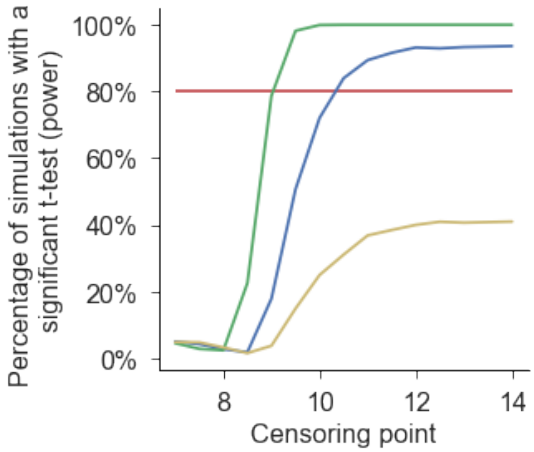


*Figure 4.7 Impact of censoring on statistical power. Population means 11.5 & 9.5 (green), 11.0 & 10.0 (blue), 10.75 & 10.25 (yellow), SDs 1.0, sample size 25, 10,000 simulations. Solid horizontal line shows target minimum power 80%.*

For the smallest difference in population means (yellow line) power peaks at around 40% even when the censoring point is so high as to have no impact. This indicates that a larger sample size is required to improve the likelihood of detecting the effect, even before considering censoring impact. The power also drops off rapidly with censoring points below 11.0, 0.25SD above the higher population mean (10.75).

For the middle difference in population means (blue line) power peaks at around 95% with no censoring, which is a high value. Power drops off rapidly with censoring points below 11.0, the higher population mean. It remains above the target 80% until the censoring point has dropped to around 10.3, which is between the two population means.

For the largest difference in population means (green line) power peaks at close to 100% with no censoring. Power drops off rapidly with censoring points below 9.5, the lower population mean. It remains above the target 80% until the censoring point has dropped to around 9.0, which is 0.5SD below the lower performing population's mean.

Considering these results together, setting the censoring point at the higher population mean ensures that power will not, on average, be reduced by censoring. This is equivalent to a maximum of 50% of the higher performance group being at ceiling.

Sample size also influences power. To investigate how this interacts with censoring and statistical power the analysis was repeated using different sample sizes (10, 25, 50 & 100) with population means of 11.0 and 10.0 (for a difference of 1.0). The results are plotted in Figure 4.8, one line per sample size.

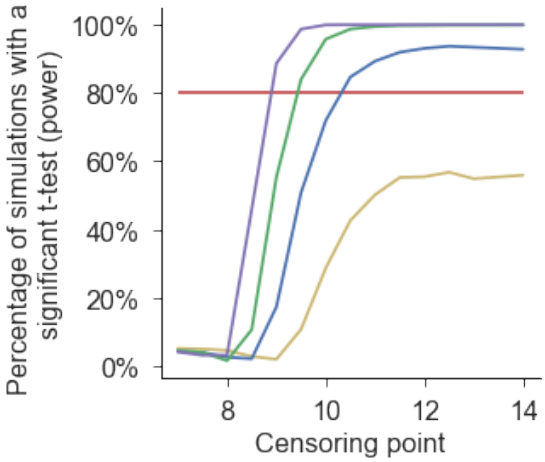


*Figure 4.8 Impact of censoring on statistical power for samples sizes of 10 (yellow), 25 (blue), 50 (green) and 100 (magenta). Population means 11.0 & 10.0, SDs 1.0, 10,000 simulations. Solid horizontal line shows target minimum power 80%.*

A sample size of 10 (yellow line) leads to a peak power of under 60% without any censoring impact, even with a large effect size (Cohen's d = 1.0 in this example). This power value is low, and again indicates that a larger sample size is more appropriate. All the other sample sizes result in good power levels when no censoring impact is present, and the higher the sample size the lower the censoring point can be before impacting power.

The key conclusions from these analyses are:

- A censoring point at or above the higher population mean ensures that power will not be reduced significantly by censoring. This is equivalent to a maximum of 50% of the higher performance group scoring at ceiling.
- Sample sizes of 10 and 25 result in inadequate power for population effect sizes of d = 1.0 and 0.5 respectively, before considering any censoring impact.
- In general, larger sample sizes increase the immunity of power to censoring.

## 4.6 Impact of censoring on the Mann-Whitney U test p-value applied to independent group means

The analysis so far has focused on the t-test. However, censoring can distort the distribution of scores such that it becomes non-normal and the normality assumption for

the t-test is no longer met. Although the t-test is robust to non-normality, there are situations in which it is more appropriate to use a non-parametric test such as the Mann-Whitney U test.

This section repeats the analysis of section 4.4, but replaces the t-test with the Mann-Whitney U test (MWU). So, the impact of censoring on MWU p-value is analysed for multiple differences between population means, and for various sample sizes. The results are illustrated below. For easy comparison, the t-test results from section 4.4 and the MWU results are positioned side by side, with the t-test on the right. The interaction of censoring and difference in population means is shown first, in Figure 4.9.

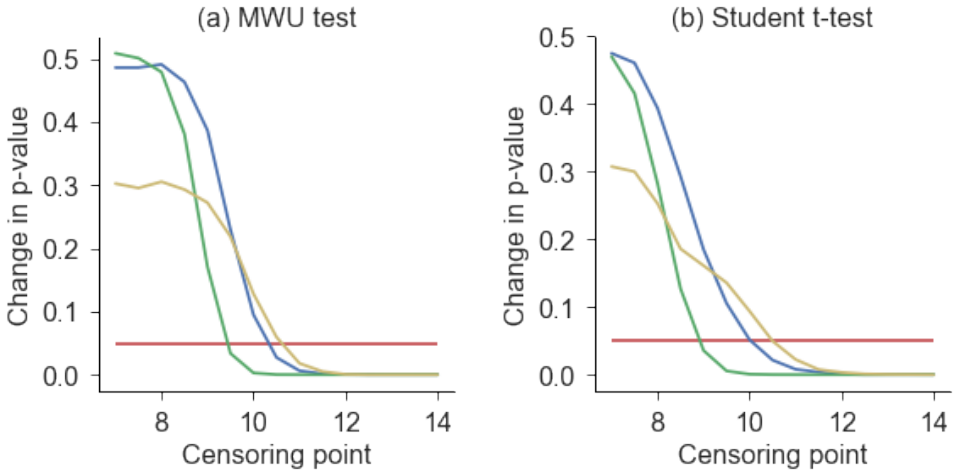


*Figure 4.9 Impact of censoring on p-values for (a) Mann-Whitney U and (b) t-tests, with varying population differences: 11.5 & 9.5 (green), 11.0 & 10.0 (blue), 10.75 & 10.25 (yellow). All SDs 1.0, sample size 25, 10,000 simulations. Red horizontal line shows an increase in p-value of .05.*

The results are very similar for both test types. However, in general the impact of censoring on MWU p-values starts at a similar censoring point level as does impact on t-test p-values, but it increases a little more rapidly as the censoring point is reduced further (steeper slope).

The interaction of censoring and sample size is shown in Figure 4.10, again with the equivalent analysis for the t-test shown for comparison.

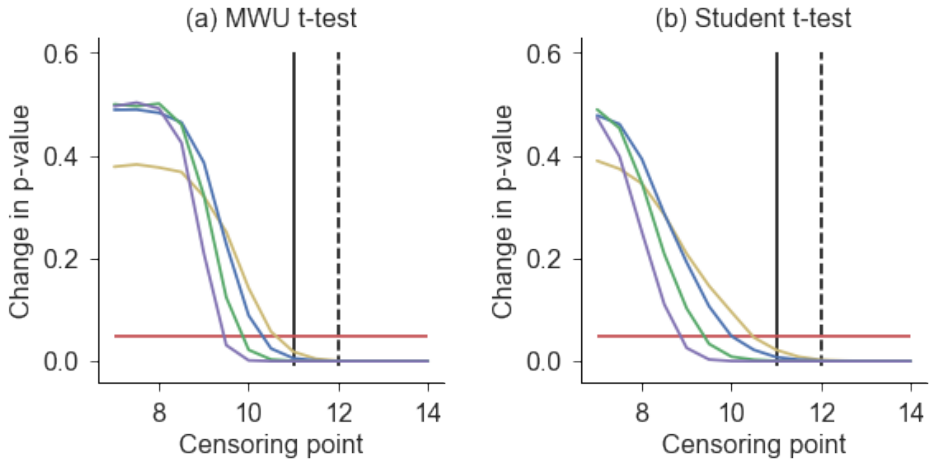


*Figure 4.10 Impact of censoring on p-values of (a) Mann-Whitney U test and (b) t-test with sample sizes of 10 (yellow), 25 (blue), 50 (green), and 100 (magenta). Population means 11.0 & 10.0, SDs 1.0, 10,000 simulations. Vertical lines show the higher population* mean (solid) and 1SD above higher population mean (dashed) respectively. Solid red horizontal line indicates an increase in p-value of 0.05.

The results are, again, very similar. However, in general the curves are a little closer together for the MWU test, indicating that while the MWU test becomes less vulnerable to censoring as sampler size increases, this benefit from larger sample size is smaller than for the t-test.

In summary, the impact of censoring on MWU test results starts at a similar censoring point as for standard t-tests, but then increases more rapidly as the censoring point is reduced further. As long as the censoring point is high enough to prevent any significant impact to t-test values (i.e. meets the guidelines identified earlier) then this will also avoid impact to MWU test results. So the same guidelines can be used for both tests.

## 4.7 Impact of censoring on forgetting rate comparisons

The reduction of a high performing group's scores at a short delay due to ceiling effects can cause this group's subsequent forgetting rate to be underestimated, since their scores at a later delay will be lower and therefore less suppressed by ceiling effects. This can cause the forgetting rate for such a high performing group to look lower than that for a lower performing group, leading to a false claim of accelerated forgetting in the lower performing group (false positive).

To investigate this quantitatively the model was extended so that two scores are simulated for each participant; one to represent a score at a short delay, and another to represent a score at a long delay. To achieve this, a fixed forgetting rate was assumed for all participants. This simulates a condition in which the forgetting rate is equal for all

participants, and therefore any statistical comparison of forgetting rates should fail to find a significant difference. Any statistical difference found would be a *false positive*.

The steps in the simulation procedure were adapted to be as follows:

1. Draw 25 samples each from two populations, with means 12.0 and 9.0, SD =1.0, for an effect size of d = 3.0. This provides the simulated scores for the short delay (t1) in the uncensored condition, for a large effect size. This very large difference between the group means was chosen as this makes the impact of censoring more visible when results are plotted; more realistic effect sizes will be used in later steps (see step 7 below).
2. Multiply all scores by a common retention factor 0.6, to simulate a common forgetting of 40% of material between the two delays for all participants. This provides the simulated scores for the long delay (t2) in the uncensored condition. In the case of population means of 12.0 and 9.0 this results in mean recall scores of approximately 7.2 and 5.4 at the long delay for the two groups.
3. Censor the samples using an initial censoring point of 14.0 to provide the short and long delay data for the censored condition
4. For both censored and uncensored datasets perform the following steps:
    a. calculate the relative forgetting rate for each participant, defined as: forgetting rate = (short delay score – long delay score) / short delay score
    b. calculate the mean forgetting rates for each group
    c. calculate the difference between the mean forgetting rates of the two groups
    d. perform a t-test comparison of the forgetting rates for the two groups, and record the p-value and the outcome of the test (significant or not significant)
5. Repeat steps 3 and 4 for multiple censoring points, between 14.0 (censoring impact should be negligible) and 5.0 (chosen to be just below the mean score for the lower performing group at the longer delay; there is no point simulating censoring below this value as almost all participants are scoring at ceiling). This gives us the impact of varying the threshold within a single simulated experiment.
6. Repeat steps 1 to 5 1000 times, to simulate running 1000 experiments.
7. Repeat steps 1 to 6 with a range of smaller effect sizes: means of 11.5 and 9.5 (d=2), 11.0 and 10.0 (d=1.0), 10.75 and 10.25 (d=0.5).

To visualise the impact of censoring on forgetting rates the forgetting curves for both groups for each censoring point are plotted in Figure 4.11. As the aim of the figure is to visually illustrate the impact of censoring, the results for the largest effect size (d = 3.0) were used as the pattern is most obvious visually in this condition.

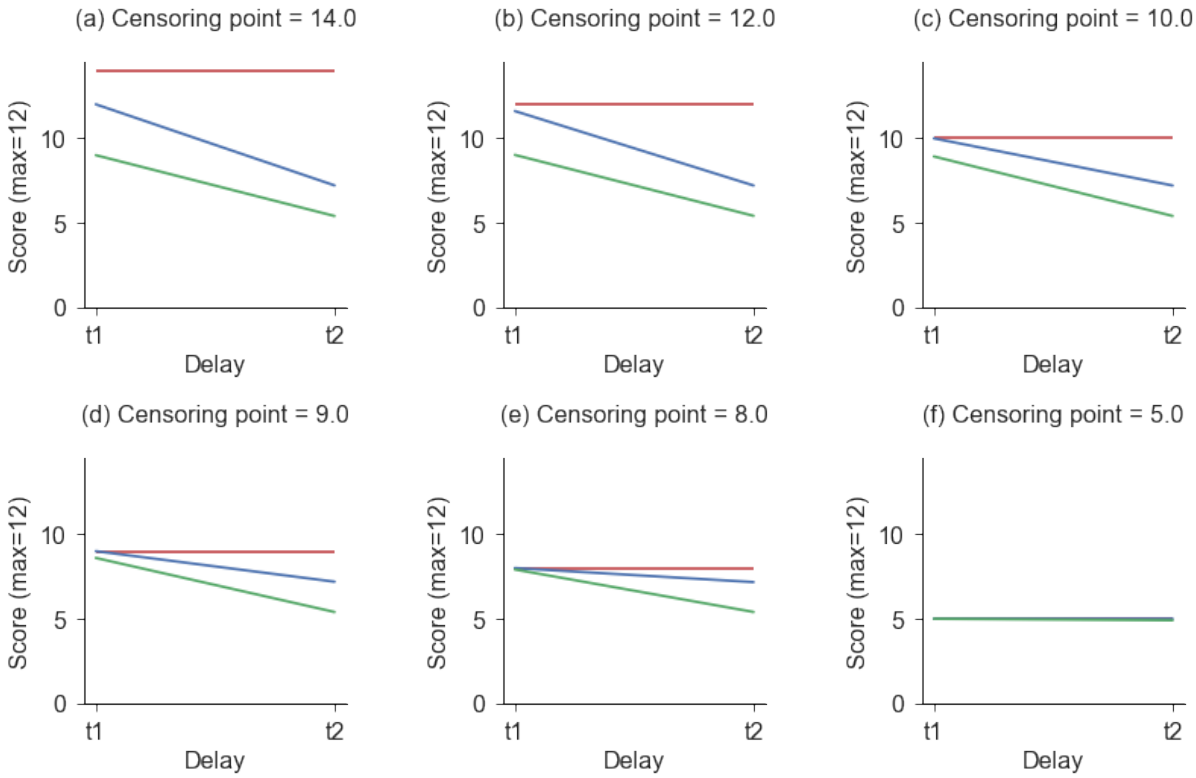


*Figure 4.11 Impact of censoring on observed forgetting rates between a short delay (t1) and long delay (t2). Population means 12.0 & 9.0, SDs 1.0, sample size 25, 1000 simulations. Horizontal red line shows the censoring point. Blue and green lines show the higher and lower performing group means.*

The relative forgetting rate formula used is the most common way to measure forgetting rates in the literature, although some studies have used absolute change in scores or other metrics. The relative forgetting approach assumes that forgetting a single item is more significant for someone who is already scoring lower, as it is a greater proportion of the items they know. For example, forgetting one item reflects a 10% loss for someone who initially recalls 10 items, while this reflects a 20% loss for someone who initially recalls 5 items. For groups performing at different levels, but with the same relative forgetting rate, the forgetting curves will not be parallel; rather they will appear to converge at longer delays. This pattern is apparent in the figure when the censoring point is set to 14.0 (Fig 4.11a). For the forgetting curves to be parallel both groups would need

to forget the same number of items, which would mean a greater relative forgetting rate for the lower performing group.

As the censoring point is lowered the slope of the forgetting curve for the higher performing group is impacted first, becoming less steep. With the censoring point set to 9.0 (Fig 4.11(d)) the apparent pattern in slopes is reversed; the slope of the higher performing groups curve (blue) is now less than that of the lower performing group (green). As the censoring point is lowered further it impacts the lower performing group too, and by the time is reaches 5.0 (Fig 4.11(f)) all apparent forgetting is eliminated, with all participants scoring at ceiling at both delays.

To investigate the impact of censoring on a t-test comparing observed forgetting rates the p-values are plotted against censoring point in Figure 4.12. For this plot a more realistic effect size of 1.0 is used (population means 11.0 & 10.0). As all simulated participants forget at the same rate, there is no difference in the pre-censoring forgetting rates of the two groups, so it is desirable that the t-test should be non-significant. However, Figure 4.12 illustrates that the mean p-value falls as the censoring point is lowered, until at approximately 10.7 the p-value becomes significant ($p < .05$), creating a false positive on average, due to censoring alone. The mean p-value remains significant until the censoring point has been further reduced to below approximately 6.2.

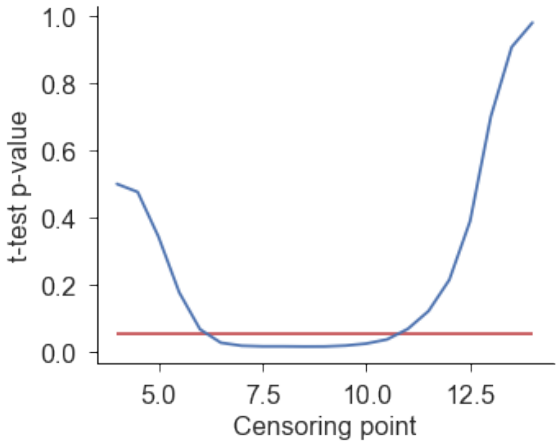


*Figure 4.12 Impact of censoring on p-value of a t-test on observed forgetting rates. Forgetting rate 40%, population means 11.0 & 10.0, SDs 1.0, sample size 25, 1000 simulations. Red horizontal line shows statistical significance level p=0.05.*

To further investigate this effect equivalent plots for a selection of group performance differences and forgetting rates are shown in Figure 4.13.

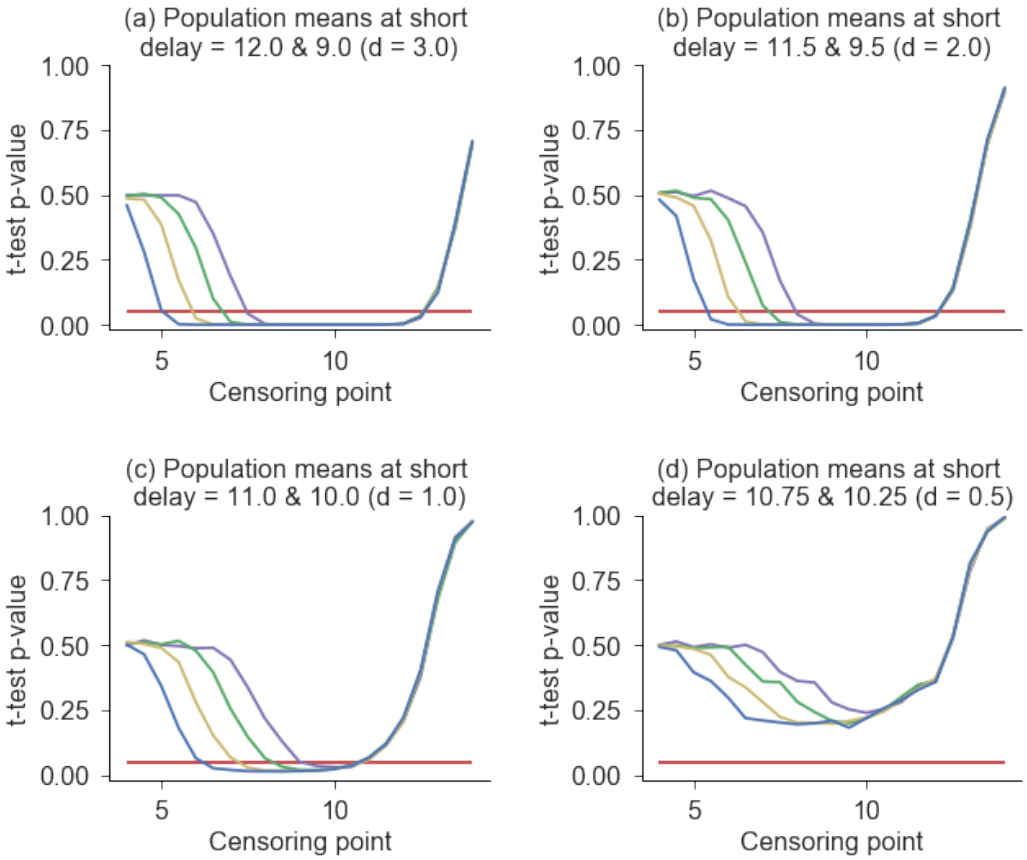


*Figure 4.13 Impact of censoring on p-value of a t-test for observed forgetting rates. Forgetting rates 10% (magenta), 20% (green), 30% (yellow), 40% (blue). Population SDs 1.0, sample size 25, 1000 simulations. Red horizontal line shows statistical significance level p = .05.*

When there is a medium difference in group performance (d = 0.5, 'medium' by convention; see Fig 4.13d) censoring alone is not enough to generate a false positive on average (across 1,000 simulations); the average p-value does not drop below .2. With larger differences in group performances (Fig 4.13a, b & c) the mean p-value can fall below .05, so censoring can be expected to generate many false positives, and the greater the performance difference the wider the range of censoring points that can generate such an error. This can be explained by reference to Figure 4.11. When there is a large difference in the performance levels between groups (large gap between the blue and green lines in Figure 4.11), the scores of the higher performing group (the blue line) can be significantly impacted by censoring before there is any impact on the lower performing group (green line); this creates the opportunity for a false-positive. By contrast, when the difference in group performance is small (small gap between the blue and green lines) both groups are likely to be impacted in a similar way, reducing the opportunity for a false-positive.

For each difference in performance, the upper censoring point where the average p-value falls below significance (p = .05) is the same for all forgetting rates; all curves cut the red line at the same point. However, the lower censoring point value at which the curves cut this line is lower for lower forgetting rates, indicating that for lower forgetting rates there is a wider band of censoring points that can trigger a false positive.

While the plots on Figure 4.13 illustrate the impact on average p-value, the more important metric for false positives is the rate at which these occur, the false positive rate (FPR). To analyse this the forgetting rate t-test significance result (significant or not significant) was used to generate plots of FPR. These are illustrated in Figure 4.14.

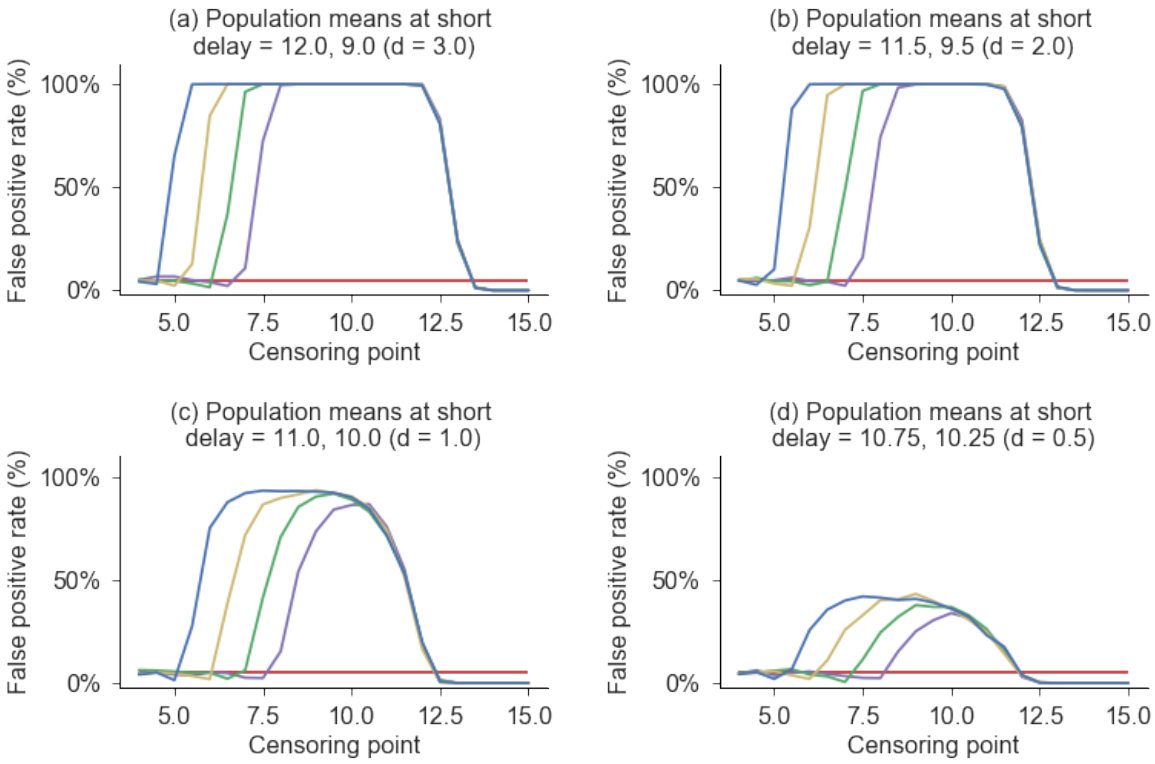


*Figure 4.14 Impact of censoring on false positive rates for t-test of forgetting rates. Forgetting rates 10% (magenta), 20% (green), 30% (yellow), 40% (blue). Population SDs 1.0, sample size 25, 1000 simulations. Red horizontal line shows an FPR of 5%.*

For all conditions the FPR starts to rise rapidly when the censoring point is at approximately 1.5SDs higher than the higher population mean score at the shorter delay. This is higher than the guidelines generated in previous sections to prevent false negatives when comparing groups at a single delay. Thus, ensuring the censoring point is at least 1.5SD higher than the higher performing group's population mean at the short delay will avoid both false negatives at single time points and false positives for forgetting rates. This figure equates to a maximum of 7% of the higher performing group scoring at ceiling at the short delay.

The peak FPR is higher for the conditions with larger differences in group performance; it peaks at ~40% for effect size d=0.5, rising to 100% for d=2.0 or larger. This again highlights that the bigger the difference in group performance the greater the opportunity for censoring to cause false positives.

Finally, to investigate the impact of sample size the analysis was repeated for the condition with effect size 1.0 at the short delay, but with sample sizes of 10, 25, 50, and 100. This is illustrated in Figure 4.15.

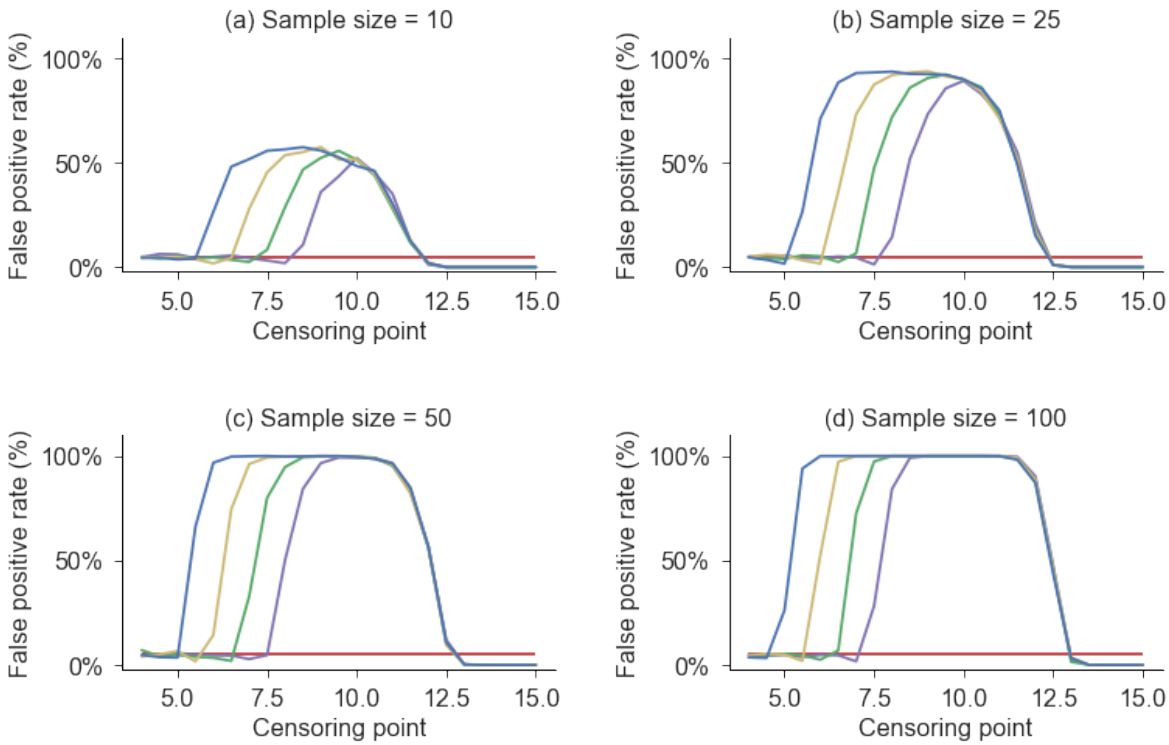


*Figure 4.15 Impact of censoring on false positive rates for t-test of forgetting rates. Forgetting rates 10% (magenta), 20% (green), 30% (yellow), 40% (blue). Population means 11.0 & 10.0, SDs 1.0, sample size (a) 10, (b) 25, (c) 50, (d) 100, 1000 simulations. Red horizontal line shows an FPR of 5%.*

Overall, figure 4.15 shows that the larger the sample size the higher the peak FPR, and the higher the censoring point that can trigger false positives. This indicates that studies with larger sample sizes are more prone to false positives due to censoring alone. Interestingly, this is the opposite of the relationship with statistical power when testing group differences at a single timepoint, where larger sample sizes improve the immunity to the effects of censoring.

The key conclusions from this analysis are:

- A censoring point at least 1.5SD higher than the higher performing group's population mean at the short delay ensures false positives will be avoided when comparing forgetting rates, assuming typical sample sizes of 25 and a population

effect size of 1.0. This figure equates to a maximum of 7% of the higher performing group scoring at ceiling.

- In general, larger sample sizes and larger group differences increase susceptibility to false positives caused by censoring when comparing forgetting rates.

## 4.8 Retrospective analysis of ceiling effects in a published ALF study

A seminal paper in the ALF literature is Butler et al. (2007). In this study, the researchers compared a group of TEA patients with a group of matched normal controls, and identified the presence of ALF between 30mins and 1 week in the TEA patients. They used a sample size of 24. Scores were presented as percentages, so the maximum score (ceiling) was 100. When testing recall for a list of words at 30 minutes the group means were 84.9 (SD 10.0) and 92.5 (SD 8.7) for the TEA and normal control groups respectively. Although it is not explicitly stated it appears they used a standard t-test, and they reported a p-value < .01 (actual value not stated).

The high score of the normal controls (92.5%) indicates many will have scored at ceiling. To analyse the impact of this the computer model used to generate the previous analyses was used in reverse. Various candidate values for population means and SDs were used as inputs, while the resulting sample means and SDs calculated by applying the ceiling were the outputs. The input values were adjusted until the output values were as close as possible to the values reported in the paper. The reported sample values and reverse engineered population values are shown in Table 4.1.

*Table 4.1 Butler et al. (2007) reported statistics at 30mins and reverse engineered population parameters.*

| Parameter | Reported values | Model output | Model population parameters |
|---|---|---|---|
| Sample size | 24 | 24 | 24 |
| ALF group mean score (SD) | 84.9 (10.0) | 84.81 (9.92) | 85.3 (11.1) |
| NC group mean score (SD) | 92.5 (8.7) | 92.52 (8.66) | 96.0 (13.1) |
| Ceiling | 100 | 100 | 100 |
| Difference in group means | 7.6 | 7.71 | 10.7 |
| Effect size (d) | 0.87 | - | 0.82 |
| t-test p-value | 0.0073 | - | 0.0038 |

This model suggests that the most likely value for the underlying difference in population means is 10.7, compared to the detected difference of 7.6, an increase of 40.8%. This illustrates how ceiling effects may have reduced the observed difference between groups at 30mins in this study.

In addition, this effect will have reduced the observed forgetting rate between 30min and 1 week for the higher performing group, and thus increased the apparent difference in forgetting rate between the higher and lower performing groups. There are too many unknowns for a reverse engineering approach to be used to evaluate quantitatively how much censoring might have distorted this forgetting rate comparison. However, using the estimated population parameters (means and SDs) the higher performing groups mean score at the 30min delay (96%) is 0.3 SDs below the maximum score (100%). With a sample size of 24, and an effect size of 0.82, behaviour would be expected to be somewhere between that indicated in Fig 4.14 (c) and (d), which suggests the FPR for a forgetting rate comparison could be as high as 30%.

While the precise numbers are very speculative, this analysis challenges Butler et al.'s conclusion that their TEA patients are only 'mildly impaired' at 30mins, and indicates there is a non-negligible risk that their claim to have detected ALF between 30min and 1 week in this group is invalid.

**NOTE**: This reverse engineering is very speculative as we only have access to summary data; we do not have the individual scores. The analysis does not provide any confidence intervals for these estimates, so these numbers must be interpreted with care. In this case, they can only be used to indicate how censoring might have impacted Butler et al.'s results. Without access to the full data set from the study we cannot take this any further.

## 4.9 Introduction to the use of Tobit analysis for analysing data where censoring is present.

While the analysis in the previous sections provides guidelines that ensure censoring does not significantly impact results obtained using standard t-tests, this leaves open the question of what to do when data does not meet the guidelines. Is there some alternative analysis that can be used when too many participants score at ceiling?

In the wider statistical literature, a technique called Tobit Regression is often used when working with censored data. This is based on a model developed by Tobin (1958). In standard linear regression, the intercept and slope parameters (regression coefficients)

are usually estimated using a technique known as Ordinary Least Squares. This approach chooses parameters that minimise the sum of the squared errors, where the error for each data-point is the difference between the observed score and the value predicted by the model.

In Tobit Regression a different technique, maximum likelihood estimation (MLE), is used. MLE chooses parameters which maximise the probability of the observed dataset occurring. For each data-point, the probability is estimated by assuming a normal distribution about the regression line; the further a point is away from the line the lower its probability of occurring. Multiplying together the probability of each data-point provides the probability for the dataset as a whole, and the regression parameters are chosen to maximise this.

The Tobit model for working with censored data uses MLE but breaks the likelihood function into two parts (McBee, 2010; Breen 1996). First, the probability of each data-point being censored is calculated across all data-points. Second, for those data-points which are not censored the probability of occurrence is estimated using standard MLE. Finally, the regression parameters are chosen to maximise both parts of the likelihood function. It has been shown that when the model's assumptions (homoscedasticity and normality of residuals) are met, the resulting parameters are unbiased estimates of the values that would be obtained using standard regression on the uncensored dataset (McBee, 2010).

When we want to analyse the difference between groups means we can use linear regression with group membership entered as a predictor variable. In the case of two groups, the group membership is coded using a single binary variable, and the regression coefficient for that variable represents the difference between the group means. If that coefficient is statistically significant then there is a statistically significant difference between the group means. This use of regression to analyse the difference in group means is a feature of the general linear model (GLM; Field, 2013; Howell, 2012). When working with censored data we can use this method with Tobit Regression to estimate the difference between the uncensored group means and hence the population means.

Tobit Regression makes the standard linear regression assumptions (homoscedasticity and normality of residuals) but it is more sensitive to violations than standard regression (Breen, 1996). These assumptions should be carefully checked when using this technique.

There is currently no Tobit Regression package or built-in routine for SPSS or JASP, but there is for R. The analysis reported here was performed using the R package *vglm*.

## 4.10 Tobit estimation of difference in population means

To investigate whether Tobit regression estimates of the difference in population means are unbiased, simulated experiments were run using this sequence of steps:

1. Draw 25 samples each from two populations, with means 11.0 and 10.0, SD =1.0
2. Censor the samples using a censoring point of 11.5, and calculate the observed difference between the means of the two censored groups. Store this value.
3. Run Tobit regression on the censored data to estimate the difference between the population means. Store this value.
4. Repeat steps 1 to 3 10,000 times, to simulate running 10,000 experiments.
5. Plot the distributions of the observed difference in means after censoring and the Tobit estimation of the difference in population means.

The two distributions are shown in Figure 4.16. The true difference in population means is 1.0, which is highlighted using a red vertical line on the plots. The observed differences after censoring are biased (Fig 4.16(a)), being centred at 0.83, but the Tobit estimates (Fig 4.16(b)) are unbiased, being centred on 1.0. Although the Tobit results are less biased, they do show greater variance and have a wider confidence interval.

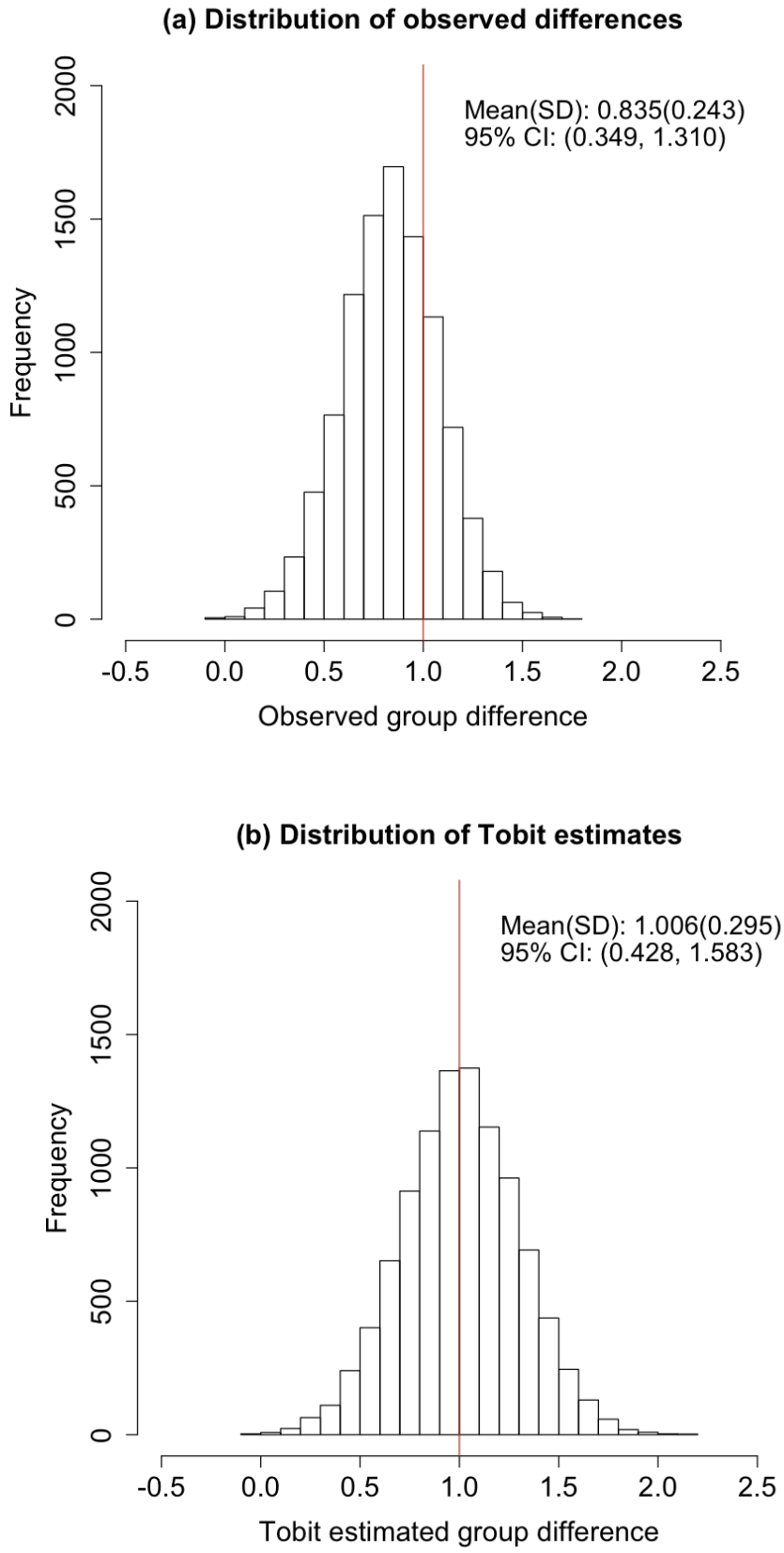


*Figure 4.16 Distribution of (a) observed difference in sample means after censoring and (b) Tobit estimates of difference in population mean. Population means 11.0 & 10.0, SDs 1.0, sample size 25, 10,000 simulations, censoring point 11.5.*

To investigate how well Tobit regression can estimate the true difference in population means for different levels of censoring the analysis was repeated for censoring point values between 9.0 and 14.0. The results are illustrated in Figure 4.17 in blue. For comparison, the difference between the means of the censored samples is shown in black, which would normally be used as an estimate of the population parameters. The true population mean difference (1.0) is shown by the red horizontal line.

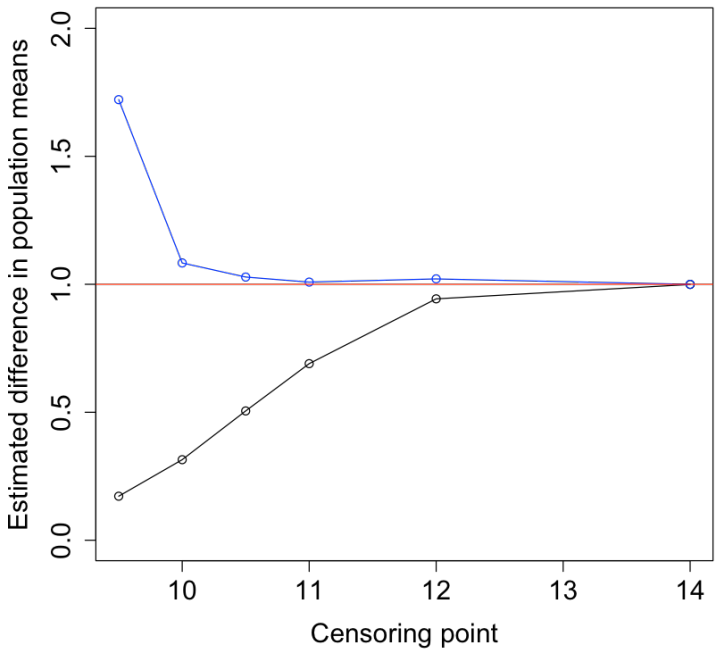


*Figure 4.17 Estimates of difference in population means generated by comparing sample means after censoring and Tobit regression. Population mean difference 1.0 (means 11.0 & 10.0), SDs 1.0, sample size 25, 10,000 simulations. Tobit estimate in blue, censored sample means in black.*

As the censoring point is reduced, the difference calculated from the censored samples falls rapidly for censoring points below 12.0. In contrast, the Tobit regression estimate tracks the true difference well until the censoring point has been reduced to approximately 10.5. This shows that Tobit regression does provide a better estimate of population differences, on average, than using the censored sample means, although it too becomes inaccurate at very low censoring points.

## 4.11 Using Tobit regression to test for significance of difference in group means

The significance of the Tobit regression coefficient provides a significance test of the difference between group means. If the coefficient is non-zero, then the group means are not equal. To analyse how effective this is, the simulation process was adjusted to also record the coefficient's significance and p-value, and the average p-value was calculated for each censoring point across all simulations. The results for average p-values are plotted in Figure 4.18, in blue. For comparison, the average value of a standard t-test comparing the censored sample means is shown in black. The standard alpha level of p = .05 is shown with a red horizontal line. The analysis was performed for population mean differences of 2.0, 1.0 and 0.5 (Figure 4.18 (a), (b), and (c)).

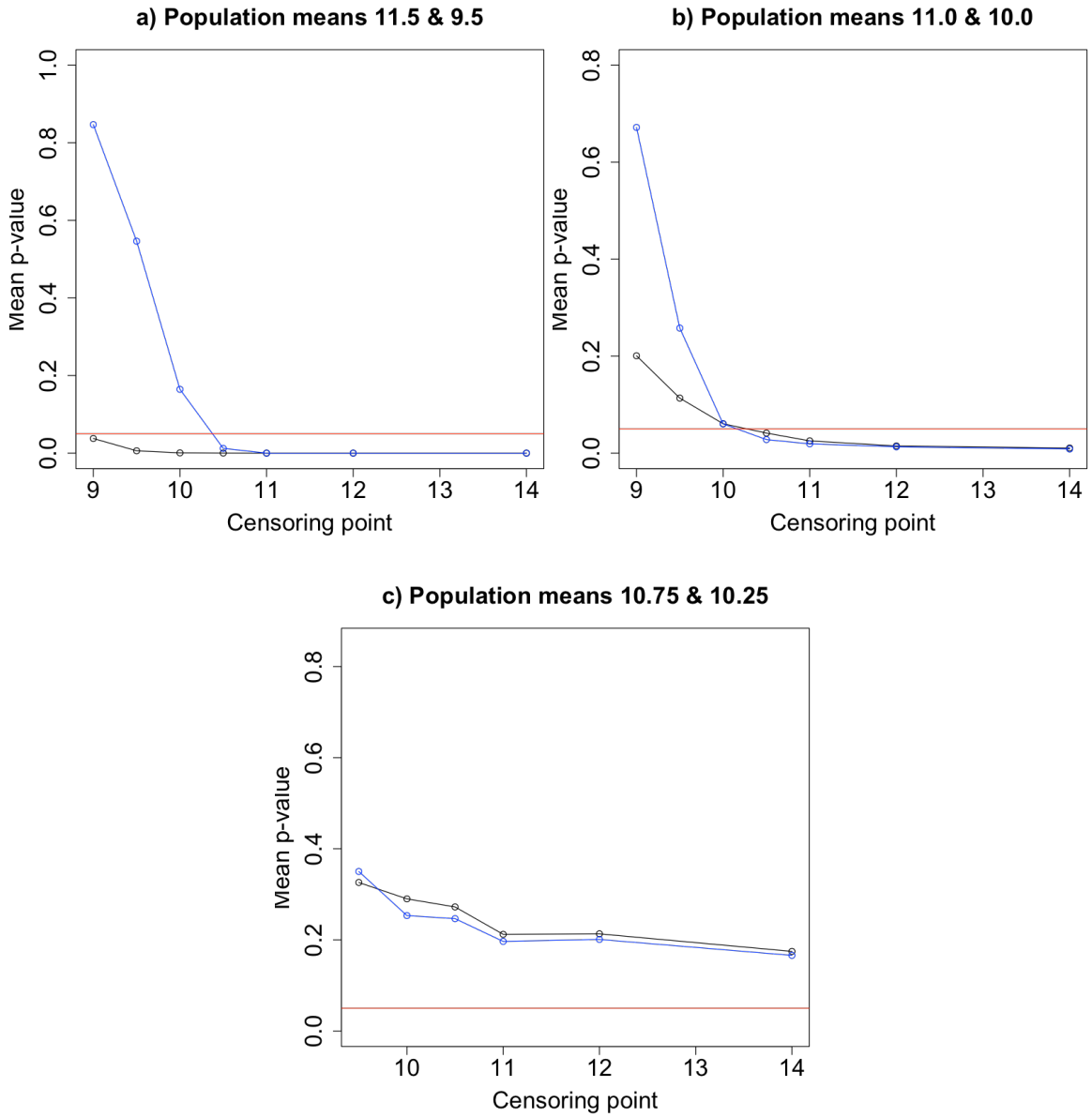


*Figure 4.18 Impact of censoring on p-value of t-test and Tobit regression comparisons of group performance. Population mean differences (a) 2.0, (b) 1.0, (c) 0.5. SDs 1.0, sample size 25, 10,000 simulations. Tobit estimate p-value in blue, t-test p-value in black. Horizontal red line indicates statistical significance level (p=0.05).*

For high censoring points there is negligible difference between the p-values obtained using Tobit regression and a standard t-test. At very low censoring point values the t-test provides lower p-values. This shows that although the Tobit regression provides a less biased estimate of the true difference in population means, it does not provide any advantage over the standard t-test when testing for statistical significance of the difference in group means.

The results of the significance test for the Tobit coefficient were used to estimate statistical power (fraction of simulations for which the coefficient was significant). This is plotted in Figure 4.19, in blue. For comparison, the equivalent power of a standard t-test of the censored sample means is shown in black. A target minimum power value of 80% is shown by the red horizontal line. The analysis was performed for population mean differences of 2.0, 1.0 and 0.5 (Figure 4.19 (a), (b) and (c)).

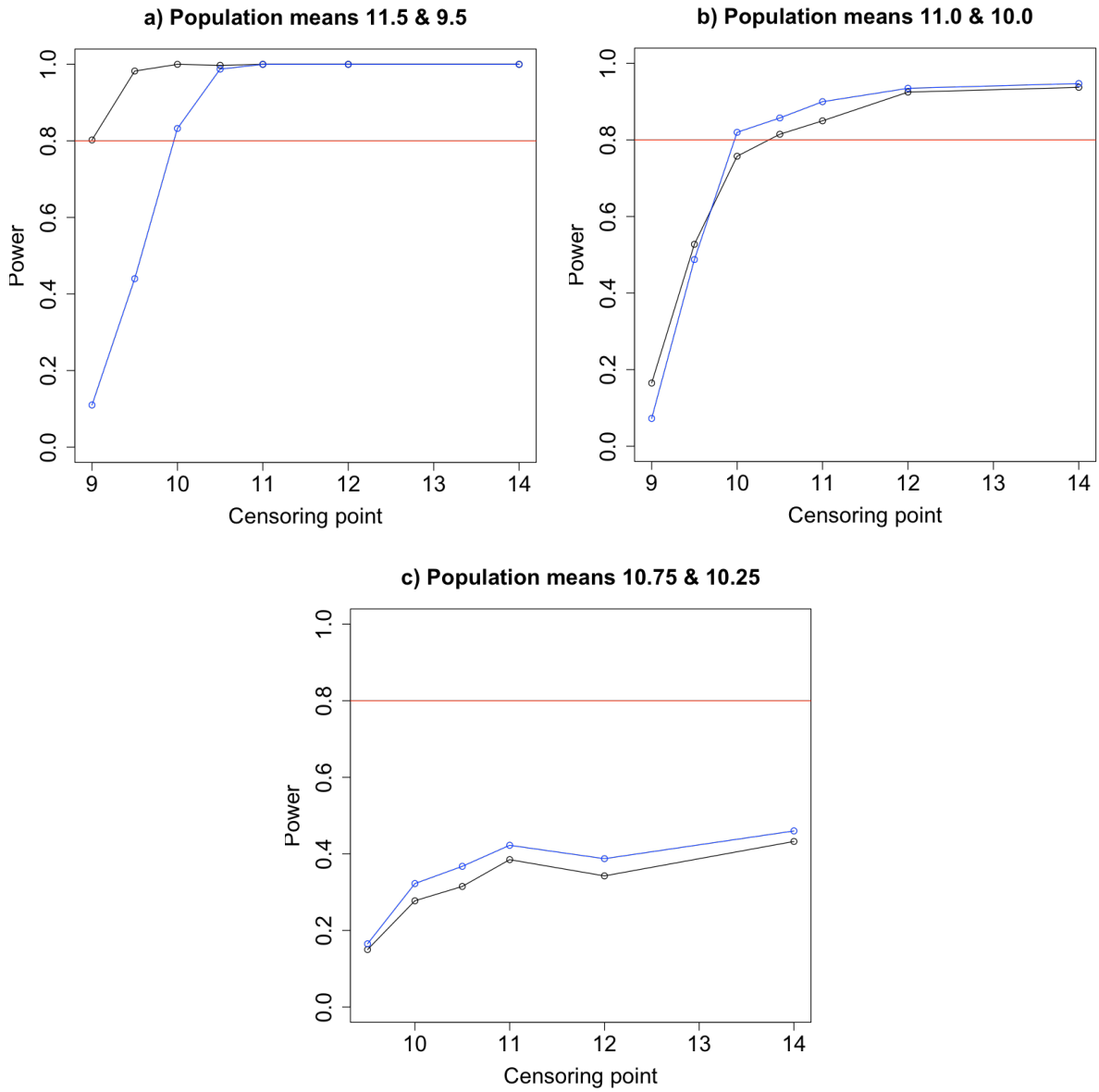


*Figure 4.19 Impact of censoring on power of t-test and Tobit regression comparisons of group performance. Population mean differences (a) 2.0, (b) 1.0, (c) 0.5. SDs 1.0, sample size 25, 10,000 simulations. Tobit power in blue, t-test power in black. Horizontal red line indicates target minimum power level of 80%.*

Overall, Tobit regression provides little if any increase in power compared to the standard t-test.

## 4.12 Using Tobit regression to test for differences in forgetting rates

Tobit regression and the associated libraries are targeted at research designs with a single censored dependent variable. This single-variable approach cannot be directly applied to forgetting rates in memory studies as it is not the forgetting rate itself which is censored; rather there are two scores which are individually censored (short and long delay) and the rate is calculated from these.

However, forgetting rates can be analysed by running Tobit regression twice, to estimate population means at both short and long delays, and then the forgetting rate for each group can be calculated from these estimates. Using this approach, forgetting rates for two groups were analysed for the case where there is no difference between groups; both groups forget at the same rate. The usual values of population means (11.0 & 10.0 at the short delay), SD (1.0) and sample size (25) were used. The estimated difference between the group forgetting rates is illustrated by the blue line in Figure 4.20. For comparison, the observed difference in forgetting rates calculated using the censored samples in shown in black, and the true difference in population forgetting rates, zero, is shown by the red horizontal line.

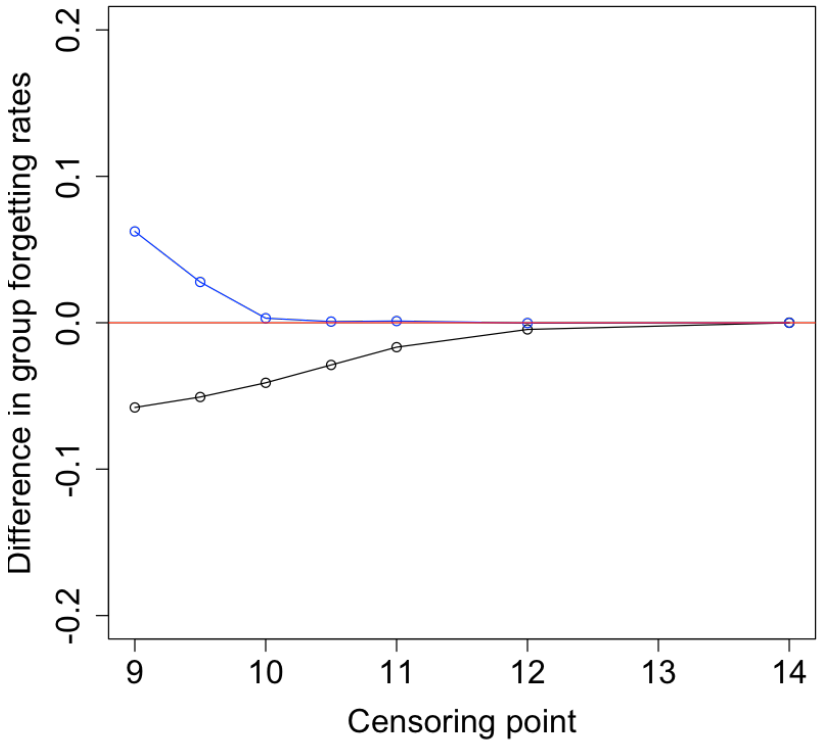


*Figure 4.20 Estimates of difference in group forgetting rates generated from sample means after censoring and from Tobit regression. Population means at short delay 11.0 and 10.0, SDs 1.0. Forgetting rate 40% for all participants (true difference is zero), sample size 25, 10,000 simulations. Tobit estimate in blue, censored sample means estimate in black.*

As the censoring point is reduced, the difference in forgetting rates calculated from the censored samples starts to diverge from zero when the censoring point is at 12.0. This can lead to false positives, claims of a difference between groups where none exists. In contrast, the value calculated by the Tobit regression procedure tracks the true population difference until the censoring point has reduced to 10.0. So, the Tobit procedure provides an estimate of forgetting rates that is less sensitive to distortion by censoring.

Although it is not performed in this analysis, confidence intervals or probability values for forgetting rate comparisons generated by this Tobit regression based procedure could be obtained from experimental data by bootstrapping. To do this, first generate multiple samples by sampling from the experimental dataset with replacement. Next, calculate the forgetting rate difference for each sample using Tobit regression. Finally, identify the values that cut off the highest and lowest 2.5% of the population of forgetting rate difference estimates to provide the confidence interval and test for significance by checking if this confidence interval includes zero.

## 4.13 Conclusions: Guidelines for minimising ceiling effects in memory research and mitigating their impact.

The analysis in this chapter confirms that ceiling effects can distort the results of statistical comparisons between groups, leading to the failure to detect a difference where one exists (*false negative*) when comparing the performance of two groups at a single delay, or the detection of a difference where none exists (*false positive*) when comparing group forgetting rates between two delays. It also provides evidence that ceiling effects have influenced the results of a seminal study in ALF research, confirming that the general concern about the possible negative impact of ceiling effects in this field is valid and not just theoretical.

It remains the case that the best way to minimise the impact of ceiling or floor effects is to design experimental tasks such that participants do not score at ceiling or floor. However, in many cases this may not be possible, and it is then necessary to identify what level of censoring can be tolerated without distorting the results of standard statistical tests such as t-tests or measurements of group differences. It is also useful to identify alternative statistical techniques which may be more appropriate when censoring is present, especially where the censoring is at a level that will distort results from standard techniques. While the following conclusions focus on ceiling effects, as those are more common, the same general approaches could be used when dealing with floor effects.

Based on the analysis in this chapter, it is possible to identify some practical guidelines for minimising and ceiling effects in memory research and mitigating their impact. When designing a study that will compare two groups at single point in time, assuming a large effect size (Cohen's d = 1.0 in this analysis), and sample size 25, the following applies:

- **Guideline 1:** to prevent censoring from distorting estimates of population means, keep the censoring point higher than 1SD above the population mean of the higher performing group. Since in practice the true population means or standard deviations are not known, this guideline can be reformulated to state that the proportion of higher performing group participants who score at ceiling must be no more than 15.9%.

Where Guideline 1 cannot be met, some distortion of population mean estimates will occur, but a statistical significance check may still produce the desired result. In this case, for the same parameters, and assuming the use of standard t-tests, these additional two guidelines apply:

- **Guideline 2:** to prevent reduction of statistical power, keep the censoring point high enough such that no more than 50% of the higher performing group are at ceiling.
- **Guideline 3:** if the censoring point is in the range between 1SD above the higher groups mean and 1SD below the lower groups mean then it is worth trying to increase the censoring point; any change will have a large impact on the observed difference between means and boost statistical power.

When designing a study that will compare forgetting rates for two groups between two time points, assuming a sample size of 25, then to avoid false positives (claiming a difference in forgetting rates where none exists) the following applies:

- **Guideline 4:** to prevent censoring from distorting estimates of population forgetting rates and leading to false positives, keep the censoring point higher than 1.5SD above the higher performing group's population mean at the short delay. Since in practice population means and standard deviations are not known this can be reformulated to state that the proportion of the higher performing group participants who score at ceiling at the short delay must be no more than 7%.

Where significant censoring is present in a dataset, Tobit regression may be an appropriate technique:

- **Guideline 5:** In the presence of censoring Tobit regression can be used to provide estimates of population means and differences in group forgetting rates that are more reliable than those generated from censored sample means. The resulting estimates will be less biased but will have slightly wider confidence intervals.

There are, of course, many experimental designs other than the simple two group comparison analysed in this chapter. It is not feasible to cover the impact of censoring on all possible designs here. However, the general approach used in this chapter can be followed to generate guidelines suited to other experimental designs as required.

# 5 Chapter 5 – Avoiding ceiling effects and optimising VALMT

## 5.1 Introduction

While experiments reported in earlier chapters have shown the usefulness of the VALMT for identifying subtle memory deficits, they have also highlighted some areas where the test could be improved. The experiments in this chapter focus on testing several possible improvements, with the overall aim of developing an optimised online VALMT.

First, previous experiments with the online VALMT have found very little forgetting between 55min and 24hr. While this might reflect very slow underlying forgetting rates over this time period, it is also possible that recalling the same pairs at both delays ('repeated recall') may be limiting forgetting through retrieval practice (Karpicke & Roediger, 2008; Roediger & Smith, 2012; Roediger & Karpicke, 2006). This highlights two issues. First, it is desirable to quantify the impact of the repeated recall on forgetting, so that experimental design decisions can be based on data rather than conjecture. Second, if it proves beneficial to remove the repeated recall then it will be necessary to learn separate sets of word-pairs to be recalled at each delay. To maintain adequate reliability and sensitivity it is desirable that both these sets be as large as possible. This raises the question of how many word-pairs can be learned in a single learning session without creating fatigue. Understanding how many pairs different age groups can learn will aid experimental design.

Second, the analysis in Chapter 4 showed that ceiling effects may have impacted the analyses of previous VALMT work and also of key published work in the relevant literature. Although existing psychology statistics texts and literature recommend avoiding ceiling and floor effects, there is no quantitative guidance on what level of these is acceptable. This is important as the nature of research into accelerated forgetting makes it very difficult to completely avoid both ceiling and floor effects when measuring memory performance at multiple delays for multiple groups of different abilities. As a result of this lack of practical guidance, two guidelines were developed which can be used to identify whether ceiling effects will impact the result of tests of statistical significance (Guideline 1: max 50% at ceiling) and the measurement of effect size (Guideline 2: max 15.9% at ceiling).

To ensure VALMT work complies with these guidelines it is necessary to find a way to reduce scores to minimise ceiling effects. While a traditional approach might involve adding more pairs to make the test harder, this approach will also increase the total time taken to learn all pairs, which may make it impossible for all participants to complete learning before they start to become fatigued or stressed. An alternative approach is to make the task more difficult by reducing the learning criterion. Requiring multiple successful recalls when learning to criterion provides retrieval practice, strengthening memories, and should increase delayed recall sores. Conversely, reducing the criterion, requiring fewer such successful recalls, should therefore lower scores and help reduce ceiling effects. Evaluating this option requires testing different learning criteria and delay pairings, to find the optimum criterion for each delay. This should be investigated for groups with different performance levels. Ideally, this information would allow a single test to be designed that measures memory performance without significant ceiling and floor effects at both delays for groups of varying memory performance levels.

For both research and clinical use, the shortest delay used for delayed recall is expected to be 30min, so the experiments in this chapter will test at 30min, thereby identifying the lowest learning criterion that will be required in practice. Even if 55min is eventually found to be the best delay for most purposes, knowing the optimum learning criteria for this shorter delay will allow VALMT to be tailored for specific purposes in future. For the longer delay these experiments retained the 24hr delay. It was expected that removing the repeated recall would reduce retention and make the test better able to detect forgetting within this period, making it unnecessary to extend this delay to several days or weeks.

It should be noted that this testing of multiple criteria is intended to find the optimum criterion for each individual delay, ensuring that the highest performing group score as high as possible without creating significant ceiling effects, leaving the maximum possible range of lower scores available for measurement of lower performing groups performance without introducing floor effects. This optimises the scoring at each delay. However, whether different learning criteria would be used in clinical practice for the two delays in an online VALMT test depends on what is to be measured. Using different criteria complicates interpretation of forgetting. Standard analysis of forgetting curves assumes the material at both delays has been learnt to the same criterion. Therefore, if it is possible to use a single learning criterion for all pairs while still avoiding both ceiling and floor effects at both delays then this would be even better. Such a design will also be

necessary if experiments show that it is not feasible to learn enough pairs to test different material at each delay. Such testing with a single criterion will be performed in the final experiment, assuming that these experiments show that adjusting the learning criterion is an effective way to control recall scores and hence reduce ceiling effects.

Third, although all VALMT versions tried so far have tested delayed recall, none has tested recognition memory. The literature review of Chapter 1 highlighted that it would be beneficial for any forgetting test to measure both recall and recognition performance. For this reason a final recognition test will be added to the VALMT, performed after the second delayed recall test to avoid disrupting any of the existing stages. Such a measure can be used to monitor recognition performance and relate this to recall performance, and investigate whether ALF impacts these capacities differentially, bearing in mind theoretical work in this area such as dual-process models (e.g. Diana et al., 2006; Yonelinas, 1994). A further possible use of a recognition test is to monitor the effort being expended by participants; are they paying attention and trying hard enough? A related question in clinical practice would be whether the participant is malingering, trying to fake having a memory problem, perhaps for insurance reasons. Since a recognition test is generally easier than a recall test, participants would be expected to score higher on recognition than recall. If anyone scores higher on recall that might indicate someone who is not paying attention or is malingering.

Finally, recruiting adequate numbers of participants for some experiments reported so far proved challenging. This reflects a wider problem in research generally. Within academia a common approach to this problem is to recruit undergraduate participants. However, this results in experimental samples which do not reflect the general population, limiting the ability to generalise findings. It is also of little use when specifically targeting other age groups, such as those over 60 years of age. Recruiting from the general population through social media is an alternative approach. While this allows recruitment of a wider range of participants, experience when running earlier VALMT experiments showed it can still take a significant amount of time to recruit the required sample sizes, especially where an experiment may be less appealing to some people due to the amount of time or effort required to participate. Another approach is to use online participant pools, where participants take part in online tasks for payment. The largest and best known of these is Amazon's Mechanical Turk service. There is evidence that Mechanical Turk can provide high quality data and samples that are more diverse than traditional samples (Buhrmester et al., 2016; Paolacci et al., 2010). It would greatly

speed up VALMT research if recruiting through this service resulted in representative samples and allowed participants to be recruited on-demand.

In this chapter four experiments are reported investigating these issues: the impact on forgetting of removing the repeated recall; the feasibility of learning two sets of stimuli in one learning session; the impact of reducing the learning criteria, and the usefulness of Mechanical Turk as a recruitment platform for VALMT research.

- Experiment 7: An initial pilot experiment testing multiple candidate learning criteria with a small group of young healthy participants recruited from Mechanical Turk. This experiment aimed to investigate whether adjusting the criteria can be used to control delayed recall levels and avoid ceiling effects at the shortest delay with a high performing group, testing with multiple candidate criteria and narrowing down these options to an optimum set. It also aimed to evaluate the feasibility of recruiting from Mechanical Turk, and investigate both the feasibility of learning two sets of stimuli in one learning session and the impact on forgetting of removing the repeated recall.

- Experiment 8: Testing the narrower set of candidate criteria identified from Experiment 7 with a larger group of young participants, recruited from first year undergraduates. This experiment aimed to validate the optimised learning criteria for 30min and 24hr recall, and further investigate the feasibility of learning two sets of stimuli in one learning session and the impact on forgetting of removing the repeated recall.

- Experiment 9: Testing the optimum criteria selected from the previous two experiments with both younger and older groups, recruited from Mechanical Turk. This experiment aimed to validate the optimised learning criteria further, ensuring both ceiling effects with younger participants and floor effects with older participants can be avoided. It also aimed to further investigate the feasibility of learning two sets of stimuli in one learning session and the impact on forgetting of removing the repeated recall.

- Experiment 10: Validating a final optimised VALMT design, developed based on the outcome of the previous three experiments. This was trialled with a large group of young participants, recruited from first year undergraduates. In addition to ensuring

both ceiling effects and floor effects are avoided, this experiment checked that key relationships between variables (e.g. the strong relationship between learning errors and subsequent delayed recall) identified in Chapter 3 are still present with the new design. This experiment also added a recognition test and evaluate this as a measure of recognition performance and as a measure for attention/effort and malingering.

## 5.2 Experiment 7: Evaluating candidate learning criteria with younger participants from Mechanical Turk

### 5.2.1 Rationale

This experiment was a small scale pilot with four aims. The first of these was to investigate the feasibility of learning enough pairs to provide separate material for recall at each VALMT delay, eliminating the need for a repeated recall at the longer delay. The first version of online VALMT required participants to learn 12 pairs. This version only provides enough for six pairs per delay if unique pairs are to be recalled at each delay; this low number would be expected to have a negative impact on reliability and sensitivity. This experiment will require participants to learn 16 pairs, providing 8 pairs to be tested at each delay.

The second aim was to investigate whether adjusting the learning criterion can provide a mechanism to control VALMT delayed recall levels and therefore adjust the risk of ceiling or floor effects. To achieve this multiple candidate learning criteria were tested with a small group of young healthy participants, with the intention that the best performing criteria would be carried forward to the next experiment for larger scale validation. This young healthy population represents the highest performing group for general research, so if optimised learning criteria can eliminate or greatly reduce the risk of ceiling effects for this group at the shortest delay, then they can be expected to eliminate ceiling effects for other groups and delays too, for example for older participants. It was identified in Experiment 6 that a 3-recall learning criterion results in a risk of ceiling effects with a young healthy population at 55min. This experiment therefore trialled 1-recall and 2-recall criteria for pairs to be recalled at the shortest delay i.e. 30min. For a repeated recall at the longer 24hr delay, Experiment 6 found a smaller, but still significant, risk of ceiling effects for pairs learnt to a 3-recall criteria. It was expected that without the repeated recall the 24hr recall scores for pairs learnt to a 3-recall criterion would be lower, and that ceiling effects would be avoided without the

need to reduce the learning criteria; to verify this a 3-recall criterion was used as one of the candidates at this delay. In fact, it was thought that without the repeated recall, scores may drop so much that lower performing groups may be at risk of floor effects. Therefore, the second criterion chosen for this delay was 5-recalls, in case scores for this longer delay needed to be increased when repeated recall was not used. This relatively high criterion will also indicate whether use of increasingly higher learning criteria leads to a form of saturation, with additional recalls during learning providing no increase in delayed recall.

The third aim was to empirically investigate the impact of removing the repeated recall. Since this experiment will include non-repeated recall testing at 24hr for pairs learnt to a 3-recall criterion, it will be possible to directly compare results with those for the Younger group from Experiment 3 which used repeated recall testing for the same delay and learning criterion. It is predicted that this will reduce scores, making any forgetting during this delay more visible.

The fourth aim was to test the feasibility of recruiting from Amazon Mechanical Turk for VALMT research. There are several potential benefits from using Mechanical Turk for psychology research, such as recruiting samples at short notice and being able to recruit from specific age groups and nationalities. There is evidence that participants recruited this way provide reliable results, and provide diverse samples that mirror the wider population (Buhrmester et al., 2016; Paolacci et al., 2010). However, there may be technical or other issues which make this approach impractical for VALMT studies. This pilot recruited a small sample from Mechanical Turk, enough to narrow down the choice of learning criteria and spot any obvious technical or procedural problems with using this service for participant recruitment.

Due to the intentionally small sample size no detailed statistical analysis was planned.

### 5.2.2 Methods

#### *5.2.2.1 Participants*

Participants were recruited through the Amazon Mechanical Turk service. They were paid £3.00 to take part. Only those who completed all three stages were paid. Mechanical Turk provides a small set of criteria which can be used to restrict recruitment to specific populations. For this experiment recruitment was limited to the 18-30yr age range, and nationality was restricted to UK and USA. Language ability is not one of the Mechanical

Turk recruitment options; limiting nationality to UK and USA was intended to work around this and ensure that the majority of participants would speak English to at least a fluent level.

As this was intended to be a pilot there was no intention to recruit a large sample, or to meet statistical power requirements. In addition, some potential participants had technical problems with completing the experiment or with receiving payment. For these reasons, recruitment was halted after 22 people had taken part.

#### 5.2.2.1.1 Inclusion criteria

To be included in the analyses a participant had to meet the following requirements, where the number in [brackets] is the number who failed to meet that criteria. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as many excluded participants people failed multiple criteria:

1. Must not report dyslexia [0]
2. Must not report a medical condition that might impact memory [0]
3. Must be aged under 30yrs [0]
4. Reported English language level must be either first language or fluent [0]
5. Must learn all 16 pairs to criterion within 20 minutes [1]
6. Must complete all 3 stages: learning, 30min & 24hr tests [7]

Mechanical Turk allows a specific age range and nationality to be specified when recruiting; specifying 18-30yrs age and nationality of USA or UK resulted in no exclusions due to age or language ability. Participants were instructed to complete the delayed tests as close to the requested time as possible (30mins, 24hrs), but to still complete the delayed tests even if they were unable to come back at the correct time. The additional inclusion criteria for an 'All_Criteria_Met' group to be compared across studies were:

7. Must complete the 30min test between 20 and 40min (30min +/- 10min) [3]
8. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [3]

As in previous experiments, the risk of exclusions distorting results was analysed by performing a group comparison between those who met all criteria (N=10) and those who met criteria 1,2,3,4 & 5 but failed to meet any of criteria 6 , 7 & 8 (either failed to

complete the delayed tests, or did these outside the acceptable time window; N=6). The two groups were compared on the number of errors made during learning to criterion. This showed that the two groups were not significantly different on errors made ($Errors_{included}$ = 20.0, $Errors_{excluded}$ = 46.0, MWU = 52.00, p=.06, r = 0.58, BF = 1.25). While this suggests excluding these participants will not have significantly impacted the results, the effect size is large and the sample size is small, so this is not conclusive. If there was an impact it will have been to raise the performance of the included group by excluding poorer performers, which would increase any ceiling effects. This should be considered when interpreting results.

#### 5.2.2.1.2 Included participants

A total of 22 individuals took part. After applying all exclusion criteria (1-8) 10 participants were included in the primary All_Criteria_Met group. A further 6 met criteria 1 to 5, but not criteria 6, 7 and/or 8 (completed learning but then either did not complete both delayed tests or failed to complete them at the correct delays), and were included in a larger 'Learnt' group to maximise data when analysing the impact of varying the learning criteria on learning performance. The demographics of these groups are summarised in Table 5.1.

*Table 5.1 Demographic information as a function of group*

| Factor | All_Criteria_Met group (N=10) | Learnt group (N=16) |
|---|---|---|
| Gender | 5M/5F | 8M/8F |
| Age Mean(SD) | 24.70(2.50) | 24.25 (2.18) |
| Education: | | |
| Batchelors | 4 | 7 |
| Diploma | 0 | 0 |
| Doctorate | 0 | 0 |
| High School | 5 | 7 |
| Masters | 1 | 1 |
| Secondary | 0 | 0 |
| Technical | 0 | 1 |
| Language: | | |
| First Language | 8 | 12 |
| Fluent | 2 | 4 |

#### *5.2.2.2 Stimuli*

For this experiment different pairs were recalled at each delay, removing the repeated recall. To provide sufficient pairs to make this possible a total of 16 pairs were learnt, instead of the usual 12, with eight to be recalled at each delay. The data from previous experiments suggested this was the maximum that could be learnt in a single learning period without creating fatigue for normal healthy young participants.

The stimuli set of 16 unrelated-word-pairs was made up of the same 12 pairs used in the previous experiments, plus four additional pairs selected following the same criteria as before, matching for familiarity, concreteness, imageability and frequency. These 16 pairs were randomly split into 4 subsets of 4 word-pairs each:

- Set 1: to be learnt to a criterion of 1 successful recall, and tested at 30min
- Set 2: to be learnt to a criterion of 2 successful recalls, and tested at 30min
- Set 3: to be learnt to a criterion of 3 successful recalls, and tested at 24hr
- Set 4: to be learnt to a criterion of 5 successful recalls, and tested at 24hr

While this only provides four pairs for each condition this was deemed sufficient as the aim was only to identify the best criteria which would then be carried forward to later experiments with only one criteria used per test delay.

For the initial presentation during learning the pairs from the four sets were interleaved. At each delayed test the pairs from the relevant two sets were interleaved.

#### *5.2.2.3 Procedure*

This experiment used the same online VALMT system and procedure as Experiments 3 to 6, with the following alterations:

- Different pairs were recalled at each delay, removing the repeated recall.
- For each delay two different learning criteria were used, with half the pairs learnt to each.
- The first delayed test was performed at 30min instead of 55min
- The MCS questions were omitted from the demographics; as the aim was to evaluate learning criteria and all participants were young and healthy, this information was deemed unnecessary

### 5.2.3 Results

#### *5.2.3.1 Mechanical Turk*

Through Mechanical Turk it was possible to recruit participants from a specific age range and nationality, which greatly simplified the recruitment process. It was also possible to recruit participants more rapidly. However, the total cost involved can be significant. Each participant was paid £3.00, and while this is a small amount to pay an individual it rapidly builds up when a large group of participants is required. This would make it impractical for large scale studies unless significant funding was available.

There were technical problems with linking the Mechanical Turk system, through which participants must be recruited and paid, to the online VALMT system. Several participants were unable to access the system to take part, and a small number experienced difficulty in requesting payment after completing all VALMT stages. Mechanical Turk is designed for tasks that can be completed in one continuous operation in a short period of time; it is not suited to tasks like VALMT which include multiple stages to be completed at different times.

Mechanical Turk policies do not allow collection of any information that might identify individuals, such as phone numbers, email addresses or names. These policies are strictly enforced. While this was not a problem for this initial VALMT research, it would be an issue for any future research that involves collecting any personal information and will also make it impossible to use automated reminders to encourage participants to return at the correct delays for later stages.

Surprisingly, the availability of a financial reward did not seem to reduce general attrition; 31% of those who completed learning failed to complete both delayed tests, compared to 32% in Experiment 4. Similarly, there was only a small difference in punctuality, with 29% of those who completed both tests completing at least one outside the acceptable window, compared to 33% for Experiment 4. Together, these results suggest that even when a small financial reward is offered, a system with automated reminders will be beneficial to reduce attrition and ensure tests are completed at the correct times.

#### *5.2.3.2 Learning performance*

The number of trials needed to complete learning is different for each criterion, even when no errors are made, so it is difficult to interpret a direct comparison on this metric.

Instead, to better allow performance to be compared, the number of errors for each criterion is summarised in Table 5.2. To maximise use of data, the analysis is performed for the larger Learnt group.

*Table 5.2 Total errors made during learning as a function of learning criterion for the Learnt group.*

| Learning Criterion | Learnt group (N=16) *Mean(SD)* |
|---|---|
| 1 recall | 6.06(4.72) |
| 2 recalls | 8.68(7.44) |
| 3 recalls | 8.00(7.38) |
| 5 recalls | 9.75(9.68) |

The standard deviations are large, as is to be expected with such a small sample and only four pairs per criterion. However, in general, the more successful recalls that are required to reach criterion the greater the number of errors made.

#### *5.2.3.3 Delayed recall performance and risk of ceiling effects*

To investigate delayed recall performance the mean recall scores for each criterion are summarised in Table 5.3, for the All_Criteria_Met group. The proportion of participants scoring at ceiling for each criterion (4 correct out of 4) is also shown.

*Table 5.3 Delayed recall performance and proportion of participants scoring at ceiling as a function of learning criterion for the All_Criteria_Met group (N=10).*

| Condition | All_Criteria_Met group (N=10) *Mean(SD)* | Proportion scoring at ceiling (4 out of 4) |
|---|---|---|
| 1-recall criterion, 30min delay | 2.90(1.10) | 30.0% |
| 2-recall criterion, 30min delay | 3.50(0.71) | 60.0% |
| 3-recall criterion, 24hr delay | 2.50(1.08) | 20.0% |
| 5-recall criterion, 24hr delay | 2.80(1.03) | 30.0% |

For each delay, learning to a higher criterion results in greater delayed recall and a larger proportion scoring at ceiling. For material to be recalled for the first time at 24hr a criterion of 3-recalls provides the best option, keeping the proportion scoring at ceiling within Guideline 1 (max 50% at ceiling) and close to Guideline 2 (max 16% at ceiling). For material to be recalled at 30min, even a 1-recall criterion still results in the proportion scoring at ceiling being above Guideline 2, although Guideline 1 is met. Since it is not

possible to use a criterion lower than 1-recall, this result suggests 1-recall is the optimum criterion to use for 30min recall, but that some ceiling effects may still be present. However, it would be expected that the proportion of participants scoring at ceiling would reduce when the number of items to be learnt and recalled using this criterion is increased to eight, as would be the case when only one criterion is used per test delay. There was also some evidence of diminishing returns for additional recalls during learning, with the increase in recall caused by raising the criterion from 1-recall to 2-recall being larger than that for raising it from 3-recall to 5-recall.

To further investigate the impact of criterion on possible ceiling or floor effects the distributions of scores are illustrated in Figure 5.1 for each criterion.

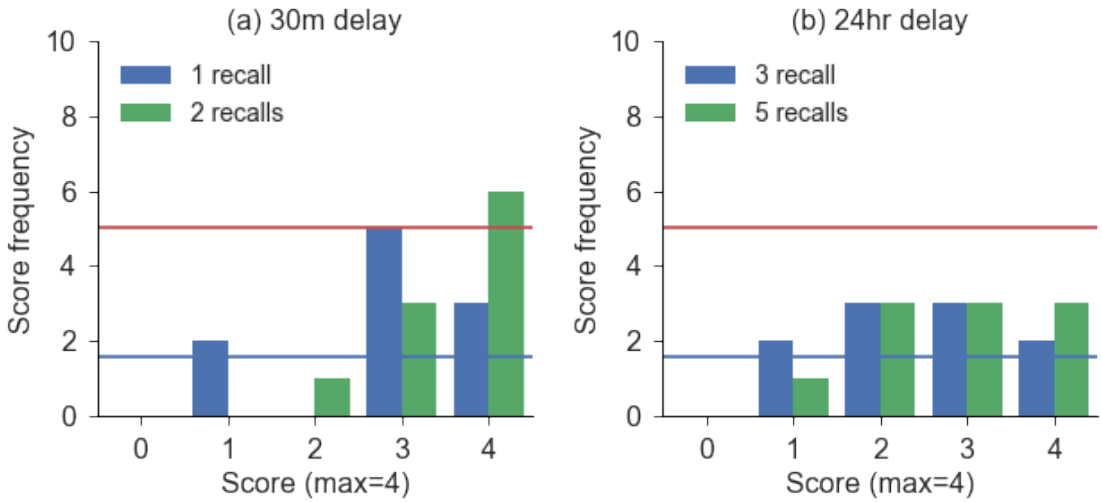


*Figure 5.1 Distribution of recall scores as function of delay and learning criterion, for the All_Criteria_Met group (N=10). Red and blue lines represent Guidelines 1 and 2 respectively.*

At both delays the lower learning criterion results in a more normal distribution, and keeps the number achieving the maximum score within, or closer to, both Guidelines. This confirms that the best criteria, which should be carried forward for larger scale validation in the next experiment, are 1-recall for 30min material, and 3-recalls for 24hr material.

#### *5.2.3.4 Learning duration when learning 16 pairs*

To investigate whether it is practical to ask participants to learn 16 word-pairs in one learning session, the total duration of learning was analysed. Figure 5.2 shows the distribution of learning durations for the All_Criteria_Met and Learnt groups.

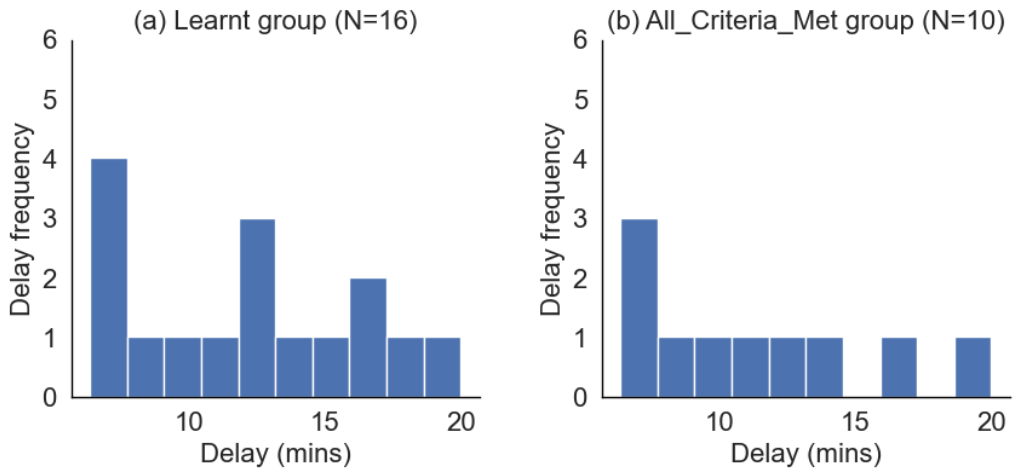


*Figure 5.2 Distribution of learning durations as function of group.*

The distribution of durations for both groups extends to 20min, with several in the 17 to 20min range. In addition, one participant was excluded for exceeding the 20min limit ('expired'). These distributions suggest that while learning 16 pairs is possible for young healthy participants, it is on the limit of what is possible without creating significant fatigue.

#### *5.2.3.5 Impact on 24hr recall of removing the repeated recall*

To investigate the impact of removing the repeated recall, the 24hr recall scores were compared with those from the Younger group from Experiment 3 where a repeated recall was used. Table 5.4 summarises the comparison.

*Table 5.4 24hr delayed recall performance for pairs learnt to a 3-recall criterion, for the All_Criteria_Met group and the Younger group from Experiment 3.*

| Variable | All_Criteria_Met group (N=10) | Expt 3 Younger group (N=49) |
|---|---|---|
| 24hr recall | 2.50/4 = 62.5% | 9.7/12 = 80.8% |

Note that the number of pairs being recalled is different in each case. For the current experiment, only four pairs were learnt to a 3-recall criterion and recalled at 24hrs. For Experiment 3, all 12 pairs were learnt to a 3-recall criterion and recalled at 24hrs. This complicates interpretation and means this comparison can only provide a tentative indication of the impact of removing the repeated recall. It does however suggest that, as expected, eliminating the repeated recall results in lower recall at the 24hr delay.

### 5.2.4 Discussion

#### *5.2.4.1 Using Mechanical Turk for recruitment*

This was the first VALMT experiment to try recruiting participants through Amazon's Mechanical Turk service, and encountered mixed results. On the positive side it was possible to recruit a sample from a targeted age range and nationality within a short period of time (approximately 1 week), which has the potential to speed up research. However, the total time taken to complete a VALMT experiment meant the amount each participant must be paid was not insignificant (£3.00), which makes the solution expensive for larger studies. There were also technical problems with linking the Mechanical Turk system and the online VALMT, and the system is not well suited to multi-stage tasks like VALMT. This resulted in an experience which was not seamless, and created issues for participants. These problems would need to be resolved before attempting to use this approach for any large scale study.

Perhaps the biggest challenge to future use of Mechanical Turk is that Amazon's policies do not allow collection of any personal identification information such as phone numbers, email addresses or names. While no such information is currently saved within VALMT, a contact detail of some sort would be required to allow for automated reminders to encourage completion of the delayed tests at the correct time. It is hard to see how this could be done without contravening Amazon's policies.

It was expected that the financial reward obtained by completing the experiment through Mechanical Turk would encourage people to complete all stages, and perhaps also result in better time-keeping, with more people completing their delayed tests at the correct times. However, neither of these things happened; the proportions of people who completed all stages and who completed both tests on time were almost identical to the equivalent proportions from Experiment 3, which recruited members of the general public through social media with no reward of any type. These results suggest that even if a small financial reward is offered in future studies, a system with automated reminders will be beneficial to reduce attrition and improve time-keeping.

#### *5.2.4.2 Optimising learning criteria to reduce risk of ceiling effects*

As predicted, it was possible to reduce delayed recall scores by reducing the number of times pairs had to be correctly recalled during learning, confirming that this provides a way to optimise the learning process in order to avoid ceiling and floor effects.

The optimum criterion for a given delay would be the one that keeps the young healthy participants score as high as possible while complying with Guidelines 1 and 2 from Chapter 4, to ensure they do not create significant ceiling effects. This then leaves the widest possible range of lower scores available for lower performing groups without them being impacted by floor effects. Based on this goal, the data from this experiment indicate that the optimum criterion for pairs to be recalled at 30min is 1-recall, while for 24hr recall a 3-recall criterion is best. These choices should now be verified using a larger sample of younger people, and then tested with older participants to verify that they do not result in floor effects for this lower performing group.

#### *5.2.4.3 Feasibility of learning 16 pairs to reduce need for repeated recall*

It is desirable that participants can learn enough pairs to support testing of different pairs at each VALMT delay, and avoid the need for a repeated recall. This will require more than the 12 pairs learnt in previous VALMT experiments, to avoid degrading the test's reliability and sensitivity. The current experiment showed that it was possible for most healthy young participants to learn 16 pairs within a 20min limit, which would provide eight unique pairs for each delay. However, many participants came close to the time limit, and one exceeded it. Evidence from previous experiments shows that older participants take longer to complete learning to criterion. Taken together, this evidence suggests many older participants would not be able to complete learning of 16 pairs within 20mins, which is then likely to lead to fatigue. If this is the case, then to be useful with this older age group, VALMT would need to use 12 pairs, and therefore it would not be possible to eliminate the repeated recall. An alternative would be to add a second learning session; however, this would make the learning process more complex and greatly extend the total time required to complete the test, which makes it less suitable for clinical use.

However, one positive factor is that using a lower learning criterion, although primarily done to control ceiling effects, will also reduce the total trials required and hence shorten the duration of learning. If a single low criterion was used for all pairs, rather than the mix of lower and higher criteria used in the current experiment, it may be possible to increase the number of word-pairs even for lower performing groups, while still staying within the 20min limit. Further testing with larger samples of both younger and older participants will be required to evaluate this.

#### *5.2.4.4 Impact of removing the repeated recall*

The aim of eliminating the repeated recall was to avoid 'retrieval practice' with its associated memory strengthening (Karpicke & Roediger, 2008; Roediger & Smith, 2012) and thereby make any forgetting more visible, making the test more sensitive. Even allowing for the small sample size and design differences between experiments, it does seem that pairs learnt to a 3-recall criterion were forgotten more rapidly in this experiment (no repeated recall) than in Experiment 3 (with repeated recall). Testing with larger sample sizes and with older participants is now required to evaluate whether this also leads to greater group differences in forgetting rates.

### 5.2.5 Conclusion

This experiment showed that Mechanical Turk could be used to recruit participants. Further use would require software development to overcome some technical issues. A further limitation is that Mechanical Turk's policies forbid collection of personal contact details, which will prevent the use of an automated reminder service for delayed tests.

It was possible to control delayed recall levels by adjusting the learning criterion, such that ceiling effects were kept within, or close to, guidelines, with the best criteria being 1-recall for pairs to be tested at 30min, and 3-recall for pairs to be recalled for the first time at 24hrs. The young healthy participants in this experiment were able to learn 16 pairs in one session; however this was at the limit of what was possible within 20mins, suggesting older groups may struggle to learn this many pairs. It also appears that eliminating the repeated recall does results in lower recall at 24hrs, as predicted. However, all of these findings must be interpreted with care, due to the small sample size. Replication with a larger sample is now required.

## 5.3 Experiment 8: Validating optimised learning criteria with an undergraduate group

### 5.3.1 Rationale

This experiment was a larger scale follow-up to Experiment 7, intended to validate the three key findings from that pilot experiment. First, that 1-recall and 3-recall criteria were optimum learning criteria for word-pairs to be recalled at 30min and 24hr respectively, when no repeated recall was used. Optimum criteria are those that result in

scores that are as high as possible, while still complying with ceiling effect Guidelines 1 & 2. Second, that removing the repeated recall leads to lower recall performance at the 24hr delay, potentially making the test more sensitive to forgetting. Finally, that learning 16 pairs in one learning session is feasible for a healthy young population, but may not be for lower performing groups such as older participants.

Due to the problems encountered using Mechanical Turk, this experiment recruited young participants from a undergraduate population. This had the added advantage that it allowed a more direct comparison of results with those from Experiment 3, which recruited from the same population. Experiment 3 used a 3-recall learning criterion for all 12 pairs, which matches the criterion used for 24hr pairs in this experiment, and had a repeated recall. The key comparisons between this experiment and Experiment 3 were how 24hr recall was impacted by removing the repeated recall, and how total learning duration was impacted by increasing the total pairs to be learnt from 12 to 16.

### 5.3.2 Methods

#### *5.3.2.1 Participants*

First year undergraduates at Goldsmiths, University of London were recruited, and completed the learning and 30m test phases ,during a lecture. Participants did not get research participation credits. As the aim was to identify whether results complied with the ceiling effect guidelines, rather than perform statistical comparisons, no a-priori power analysis was performed and no minimum sample size was targeted.

##### 5.3.2.1.1 Inclusion criteria

To be included in the analyses a participant had to meet the following requirements, where the number in [brackets] is the number who failed to meet that criteria. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as many excluded participants people failed multiple criteria:

1. Must not report dyslexia [7]
2. Must not report a medical condition that might impact memory [0]
3. Must be aged under 30yrs [4]
4. Reported English language level must be either first language or fluent [6]
5. Must learn all 16 pairs to criterion within 20 minutes [7]

6. Must complete all 3 stages: learning, 30min & 24hr tests [72]
    - Number who failed to complete 30min test: 16
    - Number who failed to complete 24hr test: 56
7. Must complete the 30min test between 20 and 40min (30min +/- 10min) [8]
8. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [17]

The risk of exclusions distorting results was analysed by performing a group comparison between those who met all criteria (N=32) and those who met criteria 1,2,3,4 & 5 but failed to meet any of criteria 6 , 7 & 8 (either failed to complete the delayed tests, or did these outside the acceptable time window; N=78). The two groups were compared on the number of errors made during learning to criterion. This showed that the two groups were not significantly different on errors made ($Errors_{included}$ = 18.5, $Errors_{excluded}$ = 26.5, MWU = 1339, p=.55, r = 0.07, BF = 0.24). The lack of significance and small effect size suggests excluding these participants will not have significantly impacted the results.

#### 5.3.2.1.2 Included participants

A total of 137 individuals took part. After applying all exclusion criteria (1-8) 32 participants were included in the primary 'All_Criteria_Met' group. A further 78 met criteria 1 to 5, but not criteria 6, 7 and/or 8 (completed learning but then either did not complete both delayed tests or failed to complete them at the correct delays), and were included in a larger 'Learnt' group to maximise data when analysing the impact on learning performance of varying the learning criteria. The demographics of these groups are summarised in Table 5.5.

*Table 5.5 Demographic information as a function of group*

| Factor | All_Criteria_Met group (N=32) | Learnt group (N=110) |
|---|---|---|
| Gender | 4M/28F | 15M/95F |
| Age Mean(SD) | 19.53(1.83) | 19.39 (2.14) |
| Education: | | |
| Batchelors | 0 | 6 |
| Diploma | 0 | 2 |
| Doctorate | 0 | 0 |
| High School | 32 | 99 |
| Masters | 0 | 0 |
| Secondary | 0 | 3 |
| Technical | 0 | 0 |
| Language: | | |
| First Language | 24 | 82 |
| Fluent | 8 | 28 |

#### *5.3.2.2 Stimuli*

The stimuli were the same 16 pairs used in Experiment 7. However, for this experiment only one learning criterion was used for the pairs to be recalled at each delay, so the pairs were treated as two sets of eight pairs:

- Set 1: 8 pairs to be learnt to a criterion of 1 successful recall, and tested at 30min
- Set 2: 8 pairs to be learnt to a criterion of 3 successful recalls, and tested at 24hr

For the initial presentation during learning the pairs from the two sets were interleaved. The interleaved order was identical to that used in Experiment 7.

#### *5.3.2.3 Procedure*

The online procedure was identical to Experiment 7, except that all eight pairs to be recalled at 30min were learnt to a 1-recall criterion, while all of those to be recalled at 24hrs were learnt to a 3-recall criterion.

For this experiment participants were not asked to take part in their own time. Instead, the experiment was run during a lecture. This meant that all participants took part at the same time, in the same environment. The learning stage was completed at the start of the lecture, and the first delayed test was completed 30mins later, also within the lecture. The

students were asked to complete the 24hr test themselves, and were told it was acceptable to be a little late or early, to allow them to work around whatever lecture schedule they had the next day.

Unlike Experiment 3, the students did not receive research participation credits.

### 5.3.3 Results

#### *5.3.3.1 Learning performance*

The number of trials needed to complete learning is different for each criterion, even when no errors are made, so it is difficult to interpret a direct comparison on this metric. Instead, to better allow performance to be compared, the number of errors for each criterion is summarised in Table 5.6. To maximise the use of data the analysis is performed for the larger Learnt group.

*Table 5.6 Total errors made during learning as a function of learning criterion for the Learnt group.*

| **Learning Criterion** | **Learnt group (N=110)** ***Mean(SD)*** |
|---|---|
| 1 recall | 13.83(10.90) |
| 3 recalls | 15.00(11.90) |

More errors were made learning pairs to a 3-recall criterion than to a 1-recall criterion. Although the effect size was small the difference was statistically significant ($Mdn_{1recall}$ = 11.0, $Mdn_{3recall}$ = 13.0, T = 1763, z = -2.49, $p$ = .01, $r$ = 0.29, BF = 2.34).

#### *5.3.3.2 Delayed recall performance and risk of ceiling effects*

To investigate delayed recall performance the mean recall scores for each delay and criterion combination are summarised in Table 5.7, for the All_Criteria_Met group. The proportion of participants scoring at ceiling for each criterion (8 correct out of 8) is also shown.

*Table 5.7 Delayed recall performance (max = 8) and proportion of participants scoring at ceiling as a function of delay and learning criterion for the All_Criteria_Met group.*

| Condition | All_Criteria_Met group (N=32) *Mean(SD)* | Proportion scoring at ceiling (8 out of 8) |
|---|---|---|
| 1 recall criterion, 30min delay | 5.66(1.82) | 21.9% |
| 3 recall criterion, 24hr delay | 6.00(1.97) | 21.9% |

For each delay, the proportion scoring at ceiling is within Guideline 1 (max 50% at ceiling) and close to Guideline 2 (max 16% at ceiling) identified in Chapter 4. This suggests the selected criteria are working as intended.

To further investigate the impact of criterion on possible ceiling or floor effects the distribution of scores are illustrated in Figure 5.3 for each delay and criterion combination.

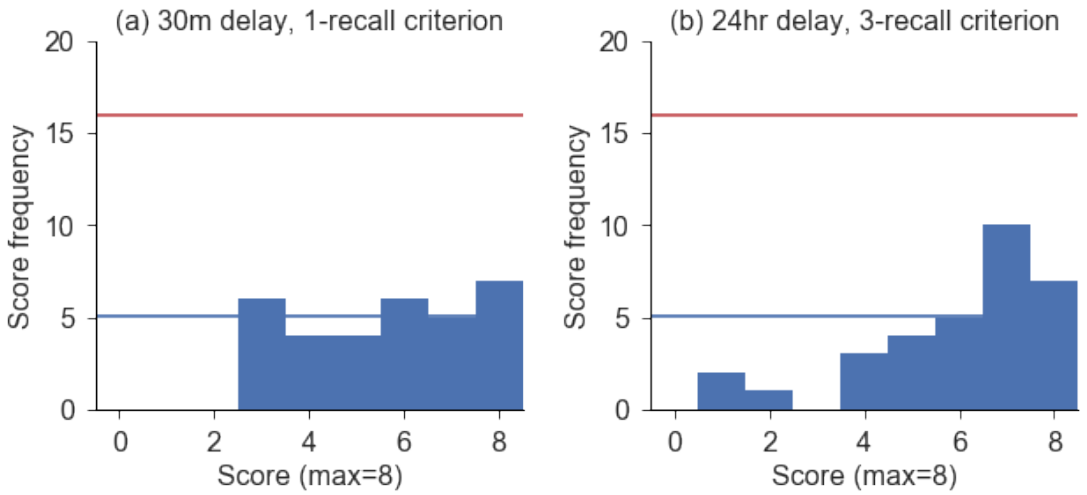


*Figure 5.3 Distribution of recall scores as function of delay and learning criterion, for the All_Criteria_Met group (N=32). Red and blue lines represent Guidelines 1 and 2 respectively.*

Inspection of the distributions in Figure 5.3 confirms that the number scoring at ceiling (max = 8) is just above Guideline 2 for both delays, which means that the scores for a young healthy group are as high as possible without creating significant ceiling effects.

The 24hr distribution (Figure 3(b)) highlights that even for these young healthy participants a small number are scoring near floor. Two participants correctly recalled only one word-pair at this delay. This suggests there is a possibility of lower performing groups such as older participants, suffering from floor effects.

#### 5.3.3.3 *Learning duration when learning 16 pairs*

To investigate whether it is practical to ask participants to learn 16 word-pairs in one learning session, the total duration of learning was analysed. Figure 5.4 shows the distribution of learning durations for the All_Criteria_Met and Learnt groups.

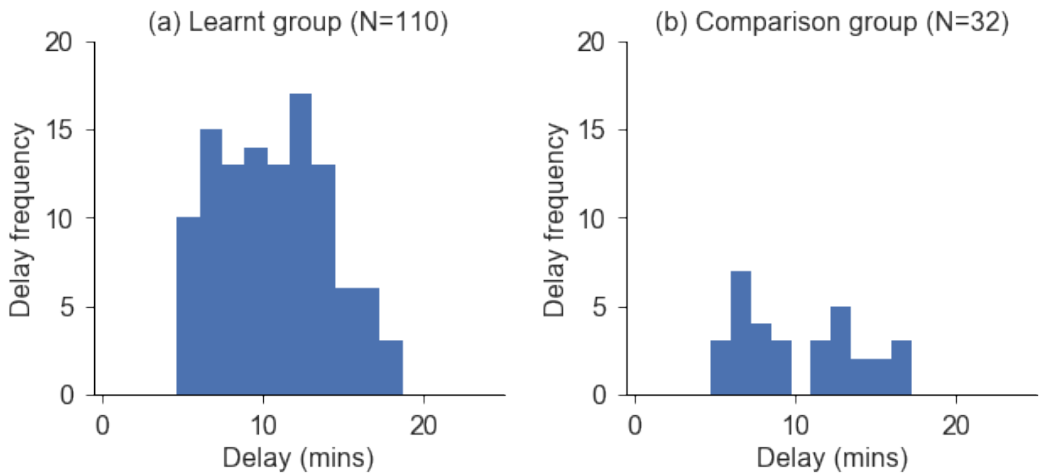


*Figure 5.4 Distribution of learning duration as function of group.*

The distribution of durations for both groups extends to, or close to, 20min. In addition, four participants who would otherwise have been included in the Learnt group were excluded for taking longer than 20mins, and similarly there was one who met all the criteria for the All_Criteria_Met group except that they took longer than 20mins. These distributions and the numbers excluded for taking longer than 20min confirm that while learning 16 pairs is possible for the majority of young healthy participants, it is on the limit of what is possible without creating significant fatigue.

#### 5.3.3.4 *Impact on 24hr recall of removing the repeated recall*

To investigate the impact of removing the repeated recall the 24hr recall scores were compared with those from the Younger group from Experiment 3, where a repeated recall was used. Table 5.8 summarises the comparison.

*Table 5.8 24hr delayed recall performance for pairs learnt to a 3-recall criterion, for the All_Criteria_Met group and the Younger group from Experiment 3.*

| Variable | All_Criteria_Met group No repeated recall (N=32) | Expt 3 Younger group Repeated recall (N=49) |
|---|---|---|
| 24hr recall | 6.00/8 = 75.0% | 9.27/12 = 77.2% |

Note that the number of pairs being recalled is different in each case. For the current experiment, eight pairs were learnt to a 3-recall criterion and recalled at 24hrs. For Experiment 3, 12 pairs were learnt to a 3-recall criterion and recalled at both 55min and 24hrs. This complicates interpretation and means this comparison can only provide a tentative indication of the impact of removing the repeated recall. The results suggest that, as expected, eliminating the repeated recall results in lower recall. However, the difference is small. A statistical comparison after converting scores to percentages showed no significant difference and a very small effect size ($t(79) = 0.44$, $p = .66$, $d = 0.10$, $BF = 0.26$).

### 5.3.4 Discussion

#### *5.3.4.1 Lecture based participation*

Running an experiment within a lecture had some advantages. It allows data from a large number of participants to be gathered in a short period of time, and provided better control of the environment as all participants completed the learning and 30m test in the same lecture hall.

However, it had the potential to add a systematic bias to the delay at which participants completed their 30m test. After all participants had completed learning the lecture continued, and at the appropriate time the lecturer asked all participants to return to the website and complete the 30m test. Since all participants returned to the website at the same time they were not able to time their delay according to the time they completed learning. As a result those who completed learning fastest completed their first delayed test at a longer delay than those who took longer to complete learning. On the other hand, as the lecturer controlled the time people returned to the website there was the potential to improve time-keeping overall, and result in more participants completing their 30m within the acceptable window (20-40min). In fact, the data showed that 30m test time-

keeping overall was better, with 92% of those who completed the 30m test doing so within the acceptable window, compared to 79% in Experiment 3 where the participant controlled their own timings. Understanding these intricacies will be useful for planning large scale data collection in future.

During Experiment 3, which also recruited from an undergraduate population, participants were rewarded for taking part by receiving research participation credits. In this experiment no such credits were awarded. This is believed to be the reason why a large dropout rate was seen at the 24hr stage. In Experiment 3, 85% of those who completed the 55m went on to complete the 24hr test. In contrast, the equivalent number for the current experiment was only 49%. This suggests that receiving research credits provides significant motivation for students to complete all stages of the experiment. This is in contrast to Experiment 8 which found that the small financial reward earned through Mechanical Turk did not result in significantly improved completion rates. Taken together, this evidence suggests the nature and significance of the reward offered is important.

#### *5.3.4.2 Optimising learning criteria to reduce risk of ceiling effects*

This experiment replicated the preliminary results from Experiment 3, and confirmed that it was possible to control the delayed recall performance at each delay by using an optimised learning criterion, titrated specifically for that delay. The selected criteria (1-recall for a 30min delay; 3-recall for a 24hr delay) resulted in the proportion of participants scoring at ceiling meeting Guideline 1 and being very close to Guideline 2. These criteria should therefore reduce ceiling effects to levels that do not significantly impact results. However, although ceiling effects will be avoided with these learning criteria, it remains possible that lower performing groups, such as older participants, may be impacted by floor effects, particularly at the 24hr delay. Future testing with older groups will be required to verify that the selected criteria achieve the best balance between preventing high performers scoring at ceiling and low performers scoring at floor.

Varying the learning criteria also impacted the number of errors made during learning, with a higher criterion leading to more errors. This makes intuitive sense, as the greater the number of successful recalls that must be made to learn to criterion, the greater the number of recall attempts made, and therefore the more opportunities there are

to make errors. Data from this experiment showed that for young healthy participants this effect is small, but statistically significant. Previous experiments have shown a strong association between learning errors and recall, with greater errors associated with poorer recall. A higher learning criterion therefore seems to create two opposing effects; the greater number of successful recalls provides retrieval practice and increases delayed recall, while making more errors will reduce delayed recall. For these young participants the retrieval practice effect appears to be dominant. It will important to review this with older participants, especially slower learners; if their learning performance is lower, the increase in error count caused by a higher learning criterion may be greater, changing the balance between these two effects.

#### *5.3.4.3 Feasibility of learning 16 pairs to reduce need for repeated recall*

Experiment 7 suggested that learning 16 pairs in one learning session was on the limit of what was feasible for young healthy participants. This experiment replicated that finding with a larger sample. In fact, even a small number of young healthy participants were unable to complete learning within the 20min limit. Since lower performing groups would be expected to require more trials to complete learning, this suggests 16 pairs may be too many for such groups and that therefore 12 pairs, as used in earlier experiments, may be the best number for a design that can be used with groups of varying performance levels. That would then mean that it will not be feasible to eliminate the repeated recall, as using only six pairs per delay will impact reliability and sensitivity. This needs to be investigated empirically by testing with older participants.

#### *5.3.4.4 Impact of removing the repeated recall*

The aim of eliminating the repeated recall was to avoid 'retrieval practice' and thereby make any forgetting more visible, making the test more sensitive. Experiment 7 suggested this might have the desired effect. However, the sample size was very small and only four of the pairs recalled at 24hr were learnt to a 3-recall criterion and could therefore be compared with Experiment 3 which used a 3-recall criterion with a repeated recall. This experiment used a larger sample size and had eight pairs learnt to a 3-recall criteria with non-repeated recall at 24hr. This provided a more reliable comparison with Experiment 3, and showed that while removing the repeated recall did result in a lower mean recall at 24hrs, the effect was very small and non-significant.

This may mean that efforts to remove the repeated recall are not needed. However, an alternative explanation would be that young healthy participants forget so little within the first 24hr that there is no opportunity for a repeated recall to make a noticeable impact. If older participants have a higher underlying forgetting rate then they may be impacted more by the removal of the repeated recall. Further testing that allows a direct age-group comparison will be required to clarify whether or not there is value in trying to remove the repeated recall.

### 5.3.5 Conclusion

This experiment showed that running a VALMT experiment within a lecture environment is feasible, and allows a large amount of data to be gathered is a short period of time. It allows the experimenter to better control the timings at which participants take the first delayed test. However, it also highlighted some intricacies around the timings which need to be born in mind when interpreting results.

Importantly, this experiment replicated key results from Experiment 7. First, recalling different word-pairs at each delay resulted in lower recall performance at 24hr, as desired; however, the effect was very small. Testing with older participants is needed to investigate whether they are impacted more by the removal of this repeated recall. Second, young healthy participants were able to learn 16 pairs within 20mins. However, this was on the limit of what they could achieve, suggesting 16 pairs will be too many for lower performing groups such as older participants, which means removing the repeated recall may not be feasible. Finally, the optimised learning criteria were validated (1-recall for 30min pairs; 3-recall for 24hr pairs), keeping scores within, or very close to, the ceiling effect guidelines at both delays. However, results also highlighted a risk that some older participants may score at floor with these criteria. Together, these results highlight the need for testing with an older age group, which should be performed in the next experiment.

## 5.4 Experiment 9: Validating optimised learning criteria with Younger and Older groups from Mechanical Turk

### 5.4.1 Rationale

Experiment 8 recruited only younger participants. This experiment was designed to replicate Experiment 8 with both younger, and most importantly, older participants. It had five key aims.

The first aim was to investigate older participants performance when using the optimised learning criteria. It was hoped that these criteria would avoid both ceiling effects for younger participants and floor effects for older participants.

The first aim was to investigate the impact on older participants of removing the repeated recall at 24hrs, and instead recalling different pairs at each delay. Experiment 8 found the impact of this design change was small for younger participants, but if older participants have a faster underlying forgetting rate then their 24hr recall performance may drop by a larger amount.

The second aim was to investigate whether it was feasible for older participants to learn 16 word-pairs in a single learning session, to provide sufficient pairs to allow different pairs to be recalled at each delay. Evidence from previous experiments suggested they may struggle to learn this many pairs within 20mins. If this is too many then removing the repeated recall may not be possible for this age group, and a test designed to cover all age groups would need to use a single set of pairs to be recalled at both delays.

The third aim was to investigate the impact of learning criteria on errors made during learning. In Experiment 8 higher learning criteria resulted in only a small increase in the number of errors made by younger participants, suggesting this age group benefit from the impact of retrieval practice with minimal negative impact from greater errors. If the number of additional errors made by older participants when learning to a higher criteria is greater, then the balance of these two opposing effects may be different for them, leading to a different pattern of results.

The final aim was to re-evaluate the use of Mechanical Turk as a source of participants. Most of the technical issues which impacted Experiment 7 had been resolved through software development, allowing this method of recruitment to be tried again.

### 5.4.2 Methods

#### *5.4.2.1 Participants*

Participants were recruited through the Mechanical Turk service. They were paid £3.00 to take part. Only those who completed all three stages were paid. For this experiment recruitment was limited to two age ranges: 18-30yr, and 60yr or over. As with Experiment 7, nationality was restricted to UK and USA to ensure that the majority of participants would speak English to at least a fluent level. As the primary aim was to identify whether results complied with the ceiling effect guidelines, rather than perform statistical comparisons between groups, no a-priori power analysis was performed and no specific minimum sample size was targeted.

##### 5.4.2.1.1 Inclusion criteria

To be included in the analyses a participant had to meet the following requirements, where the number in [brackets] is the number who failed to meet that criteria. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as many excluded participants people failed multiple criteria:

1. Must not report dyslexia [0]
2. Must not report a medical condition that might impact memory [0]
3. Must be aged under 18-30yrs or 60yrs+ [1]
4. Reported English language level must be either first language or fluent [0]
5. Must learn all 16 pairs to criterion within 20 minutes [4]
6. Must complete all 3 stages: learning, 30min & 24hr tests [10]
7. Must complete the 30min test between 20 and 40min (30min +/- 10min) [6]
8. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [6]

The risk of exclusions distorting results was analysed by performing a group comparison between those who met all criteria (N=38) and those who met criteria 1,2,3,4 & 5 but failed to meet any of criteria 6 , 7 & 8 (either failed to complete the delayed tests, or did these outside the acceptable time window; N=18). The two groups were compared on the number of errors made during learning to criterion. This showed that the two groups were not significantly different on errors made ($Errors_{included}$ = 9.5, $Errors_{excluded}$ = 16.5, MWU = 431, z = 1.56, p=.12, r = 0.26, BF = 0.75). The lack of significance and

small effect size suggests excluding these participants will not have significantly impacted the results.

#### 5.4.2.1.2 Included participants

A total of 61 individuals took part. After applying all exclusion criteria (1-8) 38 participants were included in the primary groups; 23 in a Younger group (age 18-30yrs) and 15 in an Older group (age 60yrs+). A further 18 met criteria 1 to 5, but not criteria 6, 7 and/or 8 (completed learning but then either did not complete both delayed tests or failed to complete them at the correct delays), and were included in larger Younger_Learnt and Older_Learnt groups to maximise data when analysing the impact on learning performance of varying the learning criteria. The demographics of these groups are summarised in Table 5.9.

*Table 5.9 Demographic information as a function of group*

| **Factor** | **Younger (N=23)** | **Younger_Learnt (N=35)** | **Older (N=15)** | **Older_Learnt (N=21)** |
|---|---|---|---|---|
| Gender | 13M/10F | 18M/17F | 7M/8F | 8M/13F |
| Age Mean(SD) | 24.39(1.67) | 23.74(1.87) | 65.73(4.25) | 65.90(4.45) |
| Education: | | | | |
| Batchelors | 11 | 17 | 8 | 12 |
| Diploma | 3 | 3 | 1 | 1 |
| Doctorate | 0 | 0 | 1 | 1 |
| High School | 3 | 7 | 1 | 3 |
| Masters | 3 | 4 | 3 | 3 |
| Secondary | 0 | 0 | 1 | 1 |
| Technical | 3 | 4 | 0 | 0 |
| Language: | | | | |
| First Language | 20 | 31 | 11 | 16 |
| Fluent | 3 | 4 | 4 | 5 |

The Younger and Older groups were matched on Education ($X^2(6) = 6.06$, $p = .42$, BF = 0.45), Language ($X^2(1) = 1.21$, $p = .29$, BF = 0.79) and Gender ($X^2(1) = 0.35$, $p = .55$, BF = 0.45).

#### 5.4.2.2 Stimuli

The stimuli were the same as Experiment 8: 16 pairs split into two sets of eight pairs:

- Set 1: 8 pairs to be learnt to a criterion of 1 successful recall, and tested at 30min
- Set 2: 8 pairs to be learnt to a criterion of 3 successful recalls, and tested at 24hr

For the initial presentation during learning, the pairs from the two sets were interleaved. The interleaved order was identical to that used in Experiments 7 and 8.

#### 5.4.2.3 Procedure

The online procedure was identical to Experiment 8, except that participants took part remotely and in their own time.

### 5.4.3 Results

#### 5.4.3.1 Learning performance

The number of trials needed to complete learning is different for each criterion, even when no errors are made, so it is difficult to interpret a direct comparison. Instead, to better allow performance to be compared, the number of errors for each criterion is summarised in Table 5.10. To maximise the use of data, the analysis is performed for the larger Younger_Learnt and Older_Learnt groups.

*Table 5.10 Total errors made during learning as a function of learning criterion and group.*

| **Learning Criterion** | **Younger_Learnt group (N=35)** ***Mean(SD)*** | **Older_Learnt group (N=21)** ***Mean(SD)*** |
|---|---|---|
| 1 recall | 8.54(11.36) | 9.57(10.51) |
| 3 recalls | 9.23(10.46) | 14.24(17.67) |

More errors are made learning pairs to a 3-recall criterion than to a 1-recall criterion by both groups. While for the Younger_Learnt group the effect size is small and the difference is not statistically significant ($Mdn_{1recall}$ = 3.0, $Mdn_{3recall}$ = 4.0, T = 136, z = -0.71, $p$ = .48, $r$ = 0.16, BF = 0.25), for the Older_Learnt group the effect size is very large and significant ($Mdn_{1recall}$ = 7.0, $Mdn_{3recall}$ = 8.0, T = 18.5, z = -2.56, $p$ = .01, $r$ = 0.73, BF = 13.36. This indicates that while learning to a higher criterion makes little difference to

the number of errors Younger participants make during learning, it makes a large difference to Older participants, causing them to make significantly more errors.

#### *5.4.3.2 Delayed recall performance and risk of ceiling effects*

To investigate delayed recall performance the mean recall scores for each delay and criterion combination are summarised in Table 5.11, for the Younger and Older groups. The proportion of participants scoring at ceiling (8 correct out of 8) is also shown.

*Table 5.11 Delayed recall performance (max = 8) and proportion of participants scoring at ceiling as a function of delay, learning criterion and group.*

| **Condition** | **Younger group (N=23)** | | **Older group (N=15)** | |
|---|---|---|---|---|
| | ***Mean(SD)*** | **Proportion at ceiling (8 out of 8)** | ***Mean(SD)*** | **Proportion at ceiling (8 out of 8)** |
| 1 recall criterion, 30min delay | 6.35(1.82) | 34.7% | 5.13(2.26) | 20.0% |
| 3 recall criterion, 24hr delay | 5.91(2.17) | 34.7% | 4.33(2.23) | 6.7% |

The Older group scored lower than the Younger group at both delays. Independent t-tests showed the difference was significant at 24hr ($t(36) = 2.17$, $p = .04$, $d = 0.72$, BF = 1.91), while at 30m it approached significance ($t(36) = 1.82$, $p = .07$, $d = 0.61$, BF = 1.15). The 30min result is important, as the observed effect size ($d = 0.61$) is larger than the equivalent difference between Younger and Older groups at 55min in Experiment 4 which used the 3-recall criterion ($d = 0.38$; see Section 3.3.3.3.1). This provides the first evidence that lowering the criterion to reduce ceiling effects does make the test more sensitive, as hoped.

For the Older group the proportion scoring at ceiling complies with, or is very close to, both guidelines at both delays. This suggests the selected criteria are working as desired for the Older group. However, for the Younger group the proportion scoring at ceiling is within Guideline 1 (max 50% at ceiling) but not Guideline 2 (max 16% at ceiling) at both delays. This suggests that for the Younger participants even the lowest possible criterion (1 successful recall) is not enough to guarantee compliance with both guidelines at the 30min delay.

To further investigate the impact of learning criterion on the distribution of scores and, specifically, the number of participants scoring at ceiling or floor, the distributions of scores are illustrated in Figure 5.5 for each group, delay and criterion combination.

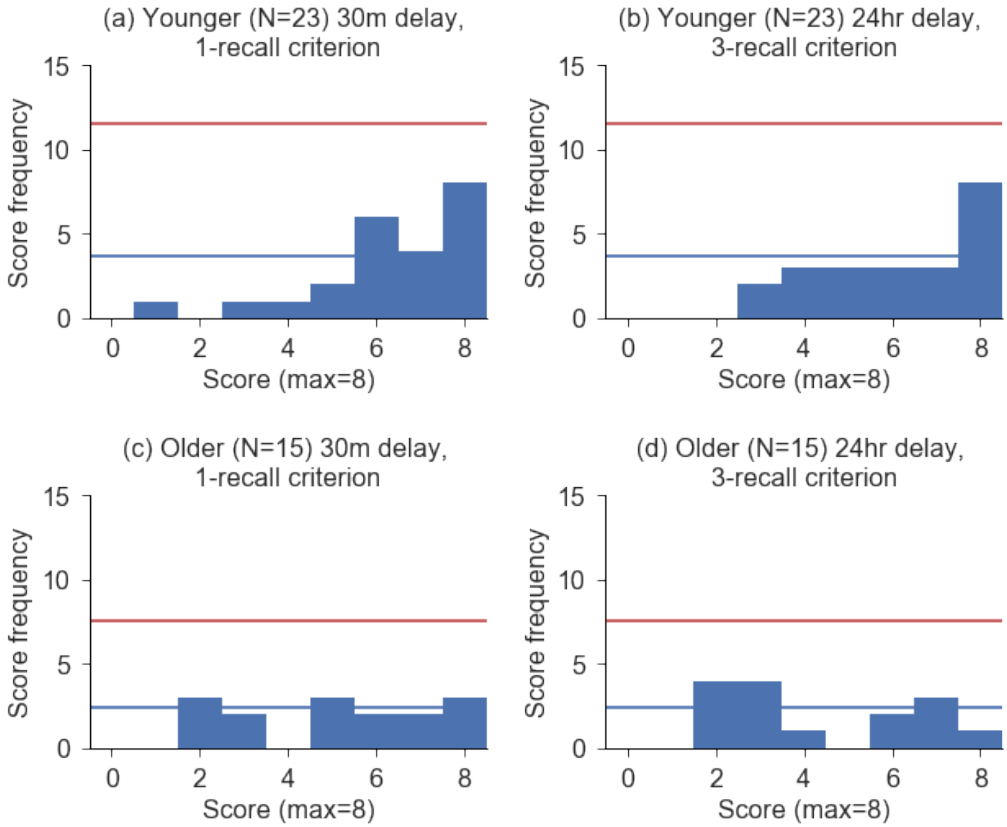


*Figure 5.5 Distribution of recall scores as function of group, delay and learning criterion. Red and blue lines represent Guidelines 1 and 2 respectively.*

Inspection of the distributions in Figure 5.5 confirms that for the Older group the number scoring at ceiling (max = 8) complies with, or is just above, both guidelines for both delays, while for the Younger group Guideline 2 (max 16% at ceiling) is met while Guideline 1 (max 50% at ceiling) is exceeded for both delays. The distributions also show no evidence of the Older group scoring so low as to cause floor effects. In summary, these results indicate that 1-recall is the best possible criterion for 30min recall for both age groups, while for 24hr non-repeated recall a 3-recall criterion is optimum for older participants but for younger participants it may be optimum to reduce the learning criterion further to 2-recall or 1-recall.

#### *5.4.3.3 Learning duration when learning 16 pairs*

To investigate whether it is practical to ask participants to learn 16 word-pairs in one learning session, the total duration of learning was analysed. Figure 5.6 shows the distribution of learning durations for each group.

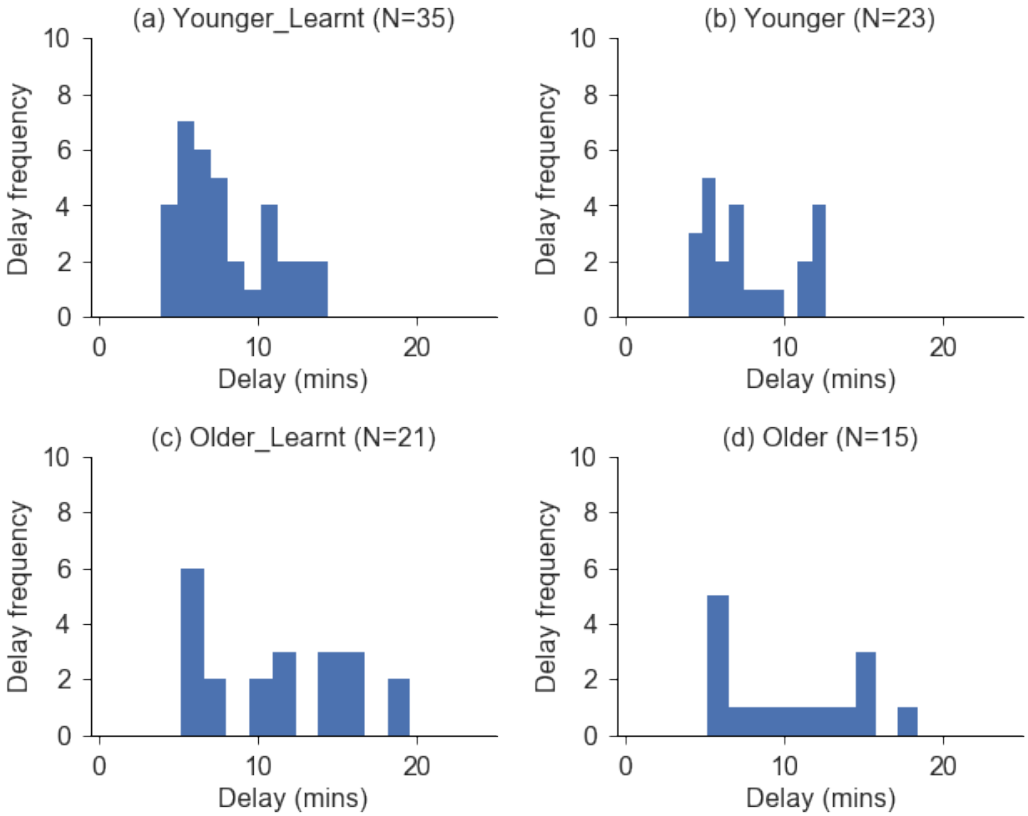


*Figure 5.6 Distribution of learning duration as function of group.*

The distribution of durations for both the Younger and Younger_Learnt groups do not extend as far as 20min, and there were no participants who were excluded for failing to complete learning within the 20min time limit. However, for both Older and Older_Learnt groups the distribution extends to, or very close to, 20min. In addition, four participants who would otherwise have been included in the Older_Learnt group were excluded for taking longer than 20mins, and similarly there were two who met all the criteria for the Older except that they took longer than 20mins. These distributions and the numbers excluded for taking longer than 20min confirm that while learning 16 pairs is possible for the majority of young healthy participants, it is beyond what is possible for some older participants without creating significant fatigue.

#### *5.4.3.4 Impact on 24hr recall of removing the repeated recall*

To investigate the impact of removing the repeated recall, the 24hr recall scores were compared with those from the Younger and Older groups from Experiment 4 which used the same learning criterion but included a repeated recall. Table 5.12 summarises the comparison.

*Table 5.12 24hr delayed recall performance for pairs learnt to a 3-recall criterion, for Younger and Older groups from the current experiment and Experiment 4.*

| Variable | Younger No repeated recall (N=23) | Expt 4 Younger Repeated recall (N=20) | Older No repeated recall (N=15) | Expt 4 Older Repeated recall (N=32) |
|---|---|---|---|---|
| 24hr recall | 5.91/8 = 73.9% | 10.6/12 = 88.7% | 4.34/8 = 54.2% | 9.19/12 = 76.6% |

Note that the number of pairs being recalled is different in each case. For the current experiment, eight pairs were learnt to a 3-recall criterion and recalled at 24hrs. For Experiment 4, 12 pairs were learnt to a 3-recall criterion and recalled at both 55min and 24hrs. This complicates interpretation and means any conclusions must be tentative.

A 2 (Age: Younger vs older) x 2 (Recall Condition: Repeated vs Non-Repeated) ANOVA showed a statistically significant main effect of both Age ($F(1, 86) = 10.76$, $p = .001$, $\eta_p^2 = 0.10$, BF = 12.33) and Recall Condition ($F(1, 86) = 14.63$, $p < .001$, $\eta_p^2 = 0.13$, BF = 52.36), but no significant interaction ($F(1, 86) = 0.60$, $p = .44$, $\eta_p^2 = 0.005$, BF = 1.28). This indicates that the Older participants score lower than the Younger participants and that, as predicted, removing the repeated recall results in lower scores. However, the lack of interaction suggests there is no significant difference in how much the different age groups benefit from the repeated recall.

Separate statistical comparisons of scores with and without repeated recall for each age group confirmed the difference is statistically significant for both Younger (t(41) = 2.17, p = .04, d = 0.67, BF = 1.90) and Older (t(45) = 3.23, p = .002, d = 1.01, BF = 15.4) participants.

#### *5.4.3.5 Comparison of participants recruited from Mechanical Turk and social media*

One factor in evaluating the use of Mechanical Turk is whether the participants recruited by this method are representative of the general population. One way to investigate this is to compare their performance levels with those recruited by other means. Table 5.13 summarises a comparison of delayed recall performance for the Younger group from the current experiment with those from Experiment 8, which recruited from a undergraduate population. The proportion of participants scoring at ceiling (8 correct out of 8) is also shown. It was not possible to perform an equivalent

comparison for the Older group as no previous experiment has used the same set of delay and criteria combinations with this age group.

*Table 5.13 Delayed recall performance (max = 8) and proportion of participants scoring at ceiling as a function of delay and learning criterion for the Younger group and equivalent group from Experiment 8.*

| Condition | Younger (N=23) | | Expt 8 Comparison (N=32) | |
|---|---|---|---|---|
| | ***Mean(SD)*** | **Proportion at ceiling (8 out of 8)** | ***Mean(SD)*** | **Proportion at ceiling (8 out of 8)** |
| 1 recall criterion, 30min delay | 6.35(1.82) | 34.7% | 5.66(1.82) | 21.9% |
| 3 recall criterion, 24hr delay | 5.91(2.17) | 34.7% | 6.00(1.97) | 21.9% |

Although the groups are matched on Language ($X^2(1) = 1.20$, $p = .27$, BF = 0.67) and Gender ($X^2(1) = 3.10$, $p = .08$, BF = 2.04), they are statistically different on Education ($X^2(6) = 43.73$ $p < .001$, BF = $5.78*10^8$). However, rather than indicating any difference in cognitive ability this may be due to the Experiment 8 group being recruited from a homogenous first year undergraduate population. For 24hr recall the difference in performance is very small and non-significant (t(53) = 0.48, p = .64, d = 0.13, BF = 0.30). For 30min recall the difference is larger, but remains non-significant (t(53) = 1.39, p = .17, d = 0.38, BF = 0.61).

The groups can also be compared on learning performance. Table 5.14 summarises a comparison of the total errors made during learning of each set of 8 pairs by the Younger group and those from Experiment 8.

*Table 5.14 Number of errors made during learning for the Younger group and equivalent group from Experiment 8.*

| Learning Criterion | Younger (N=23) *Mean(SD)* | Expt 8 Comparison (N=32) *Mean(SD)* |
|---|---|---|
| 1 recall | 8.13(11.53) | 12.84(10.20) |
| 3 recalls | 9.09(11.46) | 13.50(10.47) |

Although the Mechanical Turk recruited participants in the current experiment made less errors during learning for both sets of pairs, the differences were not statistically significant (1-recall pairs: $t(53) = 1.60$, $p = .12$, $d = 0.44$, $BF = 0.79$; 3-recall pairs: $t(53) = 1.48$, $p = .14$, $d = 0.40$, $BF = 0.68$).

Together these results suggest that young Mechanical Turk participants are similar in memory performance to participants recruited from undergraduate populations. However, the medium effect sizes seen on some comparisons, combined with the relatively modest Bayes Factors and sample sizes suggest this conclusion must remain tentative; further data would be required to confirm this equivalence.

### 5.4.4 Discussion

#### *5.4.4.1 Using Mechanical Turk for recruitment*

This was the second experiment to use Mechanical Turk as a source of participants. Recruitment during the first attempt, Experiment 7, was halted earlier than originally planned due to technical problems which meant some participants were unable to complete the test, or had problems claiming their payment. For this experiment changes were made to improve the integration of Mechanical Turk and the online VALMT. These changes were effective, with only a very small number of potential participants still having a problem. This shows that it may be technically feasible to use Mechanical Turk for recruitment. However, the wider issue of not being able to record any personal contact information means this option would still be incompatible with using automated reminders to encourage participants to return for their delayed tests at the correct time.

A further issue surrounding the use of Mechanical Turk and similar online for-payment recruiting pools is how well their participants reflect the general population. Results from the current experiment suggest that young participants from Mechanical Turk have memory performance levels similar to those from an undergraduate population. While this is far from confirming they are representative of the general population, it is at least consistent with this conclusion. Further replication and testing with larger samples and different age groups would be required to resolve this.

#### *5.4.4.2 Impact of removing the repeated recall*

Experiment 8 indicated that removing the repeated recall, and therefore the associated retrieval practice, resulted in lower recall scores at 24hr for young participants, although

the effect size was small. The current experiment replicated this general pattern, although the impact was larger, reaching a medium effect size. This difference in effect size is partly down to the choice of comparison group; the analysis for Experiment 8 used the undergraduate group from Experiment 3 so that both groups were from first year undergraduate populations, while the current experiment used the social media recruited group from Experiment 4 for the comparison group. Together, these results suggest the removal of the repeated recall does lead to lower recall for younger participants, although there is some uncertainty about the size of the effect.

The current experiment was the first to examine the impact of removing the repeated recall for older participants, and found the effect size was large for this age group. Although the difference in effect sizes between the age groups (Young: d= 0.67; Older: d = 1.01) was not statistically significant (no significant interaction) the Bayes factor suggests this evidence of no interaction is inconclusive. If removing the repeated recall does impact older participants more this might reflect a faster underlying forgetting rate in the older participants, which is exposed when the retrieval practice benefit of the repeated recall is removed. However, an alternate explanation is that younger participants performance at 24hrs even without the repeated recall may be so high that there is little opportunity for a repeated recall to raise their scores. Stronger evidence for or against a difference in how age groups are impacted would be provided by a test using a lower criterion for 24hr pairs such that scores for young participants are lower which should make any impact of the repeated recall more visible.

#### *5.4.4.3 Feasibility of learning 16 pairs to reduce need for repeated recall*

Results from Experiments 7 and 8 suggested learning 16 pairs in one session was on the limit of what was feasible for young healthy participants, and that therefore 16 pairs would be too many for lower performing groups such as older participants. The Younger group in this experiment performed a little better, although this may be down to the difference in sample size. The key result, however, is the performance of older participants. Several of the Older group were indeed unable to complete learning of 16 pairs within a 20min limit, and others came close to the limit. Since these results are for healthy older participants, the performance levels of older participants with greater memory impairments can be expected to be lower still, making 16 pairs even more of a challenge.

Taken together these findings suggest that while 16 pairs is feasible for younger healthy participants it is too many for older groups, if significant fatigue is to be avoided. This in turn suggests that for a test that can be used across multiple groups with varying memory performance levels it will be necessary to limit the learning stage to 12 pairs, as used in Experiments 1 to 6. This smaller number of pairs then leads to the conclusion that it is not practical to learn enough pairs in a single session to allow separate sets to be tested at each delay, and the repeated recall used in the online VALMT in Experiments 3 to 6 will also need to be retained. Although this potentially makes the test less sensitive at the 24hr delay, it has the advantage of increasing sensitivity at the first delay (30min or 55min) as all 12 pairs will be tested at this delay, rather than eight.

#### *5.4.4.4 Optimising learning criteria to reduce risk of ceiling effects*

This experiment replicated the results from Experiment 7 and 8, and confirmed it was possible to control the delayed recall performance at each delay by using an optimised learning criterion. The selected criteria (1-recall for 30min delay; 3-recall for a 24hr delay) resulted in the proportion of participants scoring at ceiling meeting Guideline 1 and being very close to Guideline 2 at both delays for Older participants. These criteria should therefore reduce ceiling effects to levels that do not significantly impact results. Importantly, floor effects were also avoided for this group with these criteria, suggesting the selected criteria are effective at minimising both ceiling and floor effects for this age group.

However, for Younger participants the proportion at ceiling at both delays, while meeting Guideline 1, exceeded the stricter Guideline 2. The learning criterion for 24hr recall could be reduced to 2-recall or 1-recall to further reduce scores for this group at this delay, to further improve compliance with Guideline 2. However, this option is not available for the 30min delay; since the criterion for 30min material was already set to one successful recall, it is not possible to reduce this any further while still ensuring that material has been learnt. Overall, this suggests that adjusting the learning criterion alone may not be enough to prevent ceiling effects at short delays distorting results for high performing groups such as young healthy participants, or the Super-recogniser group from Experiment 6.

The difference between the Younger and Older groups delayed recall at 30min was larger than the equivalent difference seen when using the original 3-recall criterion for

55min recall (Experiment 4). This provides early evidence that optimising the learning criterion to reduce ceiling effects may make the test more sensitive, as predicted.

Varying the learning criteria also impacted the number of errors made during learning, with a higher criterion leading to more errors, replicating the general result from Experiment 8. As in Experiment 8, the effect for Younger participants was small, indicating that varying the learning criterion makes only a small difference to the number of errors Younger participants make. However, for Older participants the impact was large, with the higher 3-recall criterion causing them to make significantly more errors. This indicates a difference in the way the learning criterion impacts different age groups. It seems that Younger participants will benefit from the retrieval practice benefit of a higher learning criterion with only a small negative impact from increased learning errors, while for Older participants the greater increase in learning errors they experience may reduce the benefit from retrieval practice.

### 5.4.5 Conclusion

This experiment showed that it was possible to eliminate most of the technical problems that disrupted integration of mechanical Turk with the online VALMT. Younger participants recruited from this service performed similarly to those recruited from an undergraduate population, suggesting Mechanical Turk samples may be representative of the wider population.

This experiment replicated three key results from Experiment 7 and 8. First, recalling different pairs at both delays resulted in lower recall performance at 24hr for both age groups, although some uncertainty remains about the size of the effect for younger participants and whether this impacts older participants more.

Second, young healthy participants were able to learn 16 pairs within 20mins. However, older participants struggled, with many failing to complete learning within 20mins, which means they may not be able to learn sufficient pairs to allow different pairs to be recalled at each delay.

Third, it was possible to control the delayed recall performance at each delay by using an optimised learning criterion; importantly, the selected criteria avoided both ceiling and floor effects for older participants at both delays. However, even with the low 1-recall criterion some younger participants scored at ceiling at 30min, suggesting it may not be possible to completely eliminate significant ceiling effects for high performing groups at short delays without more extensive changes to the VALMT paradigm.

This experiment also identified a difference in the way the learning criterion impacts different age groups; younger participants benefit from the retrieval practice arising from a higher learning criterion with only a small negative impact from increased learning errors, while older participants experience a greater increase in learning errors which may reduce the benefit they get from the additional retrieval practice.

## 5.5 Experiment 10: Validating a final optimised VALMT design including a new recognition memory task

### 5.5.1 Rationale

Based on the results of the experiments in this and previous chapters, the design for an optimised VALMT for general purpose use has become clearer. To ensure that the same test can be used across a wide range of memory abilities, while minimising ceiling and floor effects as much as possible, and avoiding fatigue effects for lower performing groups, VALMT should use a single set of 12 word-pairs, learnt to a 1-recall criterion, with all pairs recalled at 55min and at 24hrs. This final design was validated in this experiment. A group of healthy young participants were recruited from an undergraduate population, as they represent a high performing group, and hence will allow the presence of ceiling effects to be most easily detected. Assuming ceiling effects are successfully avoided for this group, then future work can validate the optimised design with older participants and other lower performing groups, to verify that floor effects are avoided.

This experiment also added a recognition memory test. This was added after the second delayed recall test, at 24hrs delay, so that it would not disrupt any of the previous stages. Recognition tests can be implemented in many ways. For this experiment a forced multiple choice design was used. In this type of test the participant is provided with multiple possible answers, and must choose one answer. One of these answer options is correct (‘target’) while the others are incorrect (‘foils’). The greater the number of answer options, the lower the score that would be obtained by random guessing. For example, with a four option test the base rate achievable by chance through random guessing is 25%. Any observed score must be interpreted based on this chance rate. It is also possible to use different types of foils. These can be completely new stimuli which have not been seen before. Selecting one of these as an answer would indicate that the participant was guessing randomly. Other foils can be stimuli which have been seen before, although they are not correct for the given question. Selecting one of these might indicate that the

participant is relying on a sense of familiarity. By using a mixture of these types of stimuli it is possible to get a more nuanced result than a simple correct/incorrect result. For the VALMT recognition test this mixed approach was used.

Analyses in earlier chapters identified some key relationships between VALMT variables. Most importantly, a strong relationship was found between errors made during learning and subsequent delayed recall. Participants who make the most errors tend to score lowest, and word-pairs which generate most errors are recalled most poorly. These findings were based on the original 3-recall learning criterion design. Since the optimised VALMT design uses a 1-recall criterion, it is important to understand whether these key relationships still apply. This experiment repeated the relevant analyses, comparing results against previous data.

### 5.5.2 Methods

#### *5.5.2.1 Participants*

First year undergraduates were recruited through the Goldsmiths, University of London research participation scheme. Participants were only awarded participation credits if they completed all stages of the experiment.

For the planned statistical comparisons with results from Experiment 3, which tested undergraduate using a 3-recall learning criteria (N=49), power analysis indicated a minimum sample size of 52 was required in order to detect at least a medium effect.

##### 5.5.2.1.1 Inclusion criteria

To be included in the analyses a participant had to meet the following requirements, where the number in [brackets] is the number who failed to meet that criteria. Note that the number of participants excluded is less than the sum of the numbers in [brackets] as many excluded participants people failed multiple criteria:

1. Must not report dyslexia [13]
2. Must not report a medical condition that might impact memory [7]
3. Must be aged under 30yrs [8]
4. Reported English language level must be either first language or fluent [4]
5. Must learn all 12 pairs to criterion within 20 minutes [0]
6. Must complete all 3 stages: learning, 55min & 24hr tests [17]

Participants were instructed to complete the delayed tests as close to the requested time as possible (55mins, 24hrs), but to still complete the delayed tests even if they were unable to come back at the correct time. The additional inclusion criteria for a 'All_Criteria_Met' group to be compared across studies were:

7. Must complete the 55min test between 45 and 65min (55min +/- 10min) [17]
8. Must complete the 24hr test between 20 and 28hr (24hr +/- 4hr) [18]

As in previous experiments, the risk of exclusions distorting results was analysed by performing a group comparison between those who met all criteria (N=52) and those who met criteria 1,2,3,4 & 5 but failed to meet any of criteria 6 , 7 & 8 (either failed to complete the delayed tests, or did these outside the acceptable time window; N=32). The two groups were compared on the number of errors made during learning to criterion. This showed that the two groups were significantly different on errors made ($Errors_{included}$ = 10.5, $Errors_{excluded}$ = 23.0, MWU = 1081, p=.02, r = 0.30, BF = 1.47). This suggests excluding these participants has the potential to have impacted the results. If there was an impact it will have been to raise the performance of the included group by excluding poorer performers, which would increase any ceiling effects. This should be considered when interpreting results.

#### 5.5.2.1.2 Included participants

A total of 122 individuals took part. After applying all exclusion criteria (1-8) 52 participants were included in the primary All_Criteria_Met group. A further 32 met criteria 1 to 5, but not criteria 6, 7 and/or 8 (completed learning but then either did not complete both delayed tests or failed to complete them at the correct delays), and were included in a larger 'Learnt' group to maximise data when analysing the impact of varying the learning criteria on learning performance. The demographics of these groups are summarised in Table 5.15. To facilitate comparisons with the data from Experiment 3, which used the same design except with a 3-recall criterion, the demographics for the equivalent groups from that experiment are also shown.

*Table 5.15 Demographic information as a function of group*

| Factor | All_Criteria_Met group (N=52) | Learnt group (N=84) | Expt 3 All_Criteria_Met group (N=49) | Expt3 Learnt group (N=84) |
|---|---|---|---|---|
| Gender | 5M/46F/1Other | 8M/75F/1Other | 11M/38F/0Other | 14M/70F/0Other |
| Age Mean(SD) | 19.15(2.33) | 19.35(2.39) | 19.33(1.85) | 19.40(2.35) |
| Education: | | | | |
| Batchelors | 2 | 4 | 0 | 1 |
| Diploma | 2 | 6 | 3 | 4 |
| Doctorate | 0 | 0 | 0 | 0 |
| High School | 44 | 64 | 45 | 78 |
| Masters | 0 | 1 | 0 | 0 |
| Secondary | 3 | 7 | 1 | 1 |
| Technical | 1 | 2 | 0 | 0 |
| Language: | | | | |
| First | 37 | 61 | 33 | 61 |
| Fluent | 15 | 23 | 16 | 23 |
| Total MCS | 2.96(2.47) | 3.00(2.30) | 2.33(2.34) | 2.80(2.69) |

The All_Criteria_Met groups from the current study and Experiment 3 are matched on Gender ($X^2(2) = 3.93$, $p = .14$, BF = 1.35), Age (t(99) = 0.41, p = .68, d = .08, BF = 0.23), Education($X^2(6) = 4.13$, $p = .39$, BF = 0.58), Language($X^2(1) = 0.17$, $p = .68$, BF = 0.28) and Total MCS score (t(99) = 1.32, p = 0.19, d = 0.26, BF = 0.46),

The Learnt groups from the current study and Experiment 3 are matched on Gender ($X^2(2) = 2.81$, $p = .25$, BF = 0.67), Age (t(166) = 0.16, p = 0.87, d = 0.03, BF = 0.17), Language($X^2(1) = 0.00$, $p = 1.0$, BF = 0.21) and Total MCS score (t(166) = 0.52, p = 0.60, d = 0.08, BF = 0.19). Although the difference in Education approaches significance ($X^2(6) = 11.08$, $p = .05$, BF = 4.84) inspection of the numbers per education level shows the difference is small, and as these are two samples from first year undergraduates at the same university it is very unlikely that this small difference represents a significant difference in cognitive ability.

#### *5.5.2.2 Stimuli*

For the learning and delayed recall testing stages this experiment used the same 12 word-pairs as Experiments 2 to 6.

For the recognition test three foils were prepared for each word-pair; along with the correct answer (target) this provided four options for the participant to choose from. For each pair one of the foils was the second word from another pair they had learnt, with each second word used once in this manner. Therefore, during the recognition test the second word in each word-pair was presented once as a target (correct answer) and once as a foil for another word-pair. The new previously unseen foils were generated in the same manner as the words for the original pairs, matching for familiarity, concreteness, imageability and frequency. They were allocated as foils to word-pairs randomly.

The two types of foils allowed two types of error to be identified. First, the intrusion of a previously seen word, which might reflect a response based on familiarity. Second, choosing a word which had not been seen before, which might indicate guessing.

#### *5.5.2.3 Procedure*

This experiment used the same online VALMT system and procedure as Experiments 3 to 6, with the following alterations:

- All pairs were learnt to a 1-recall criterion, rather than 3-recall
- A recognition test was performed after the second delayed recall test

For the recognition test, the participants were presented with the first word in a pair along with four options for the matching second word. The participant had to choose one of the four options; they were asked to guess if they did not know the answer (four option forced choice recognition).

Participants took part at a time that was convenient for them. They were asked to do this in a quiet environment where they would not be disturbed.

### 5.5.3 Results

#### *5.5.3.1 Learning performance*

The number of trials needed to complete learning is different for different learning criteria, even when no errors are made, so it is difficult to interpret a direct comparison on this metric (trials to complete learning) across different criteria. Instead, to better allow performance to be compared across criteria, the number of errors made during learning for the current experiment (12 pairs learnt to a 1-recall criterion) and Experiment 3 (12 pairs learnt to a 3-recall criterion) are summarised in Table 5.16.

*Table 5.16 Total errors made during learning as a function of learning criterion for the All_Criteria_Met group.*

| Learning Criterion | All_Criteria_Met group (N=52) *Mean(SD)* | Experiment 3 All_Criteria_Met group (N=49) *Mean(SD)* |
|---|---|---|
| 1 recall | 14.31(13.11) | - |
| 3 recall | - | 21.96(18.17) |

Significantly fewer errors were made learning to a 1-recall criterion than a 3-recall criterion, although the effect size was small, which aligns with evidence from earlier experiments that for young participants learning to a higher criterion does not lead to a large number of additional errors (Mdn1recall = 10.5, Mdn3recall = 20.0, $MWU = 960$, $p = .03$, $r = 0.25$, BF = .99).

### *5.5.3.2 Delayed recall performance and risk of ceiling effects*

To investigate delayed recall performance and the presence of ceiling effects the mean recall scores and the proportion of participants scoring at ceiling (12 correct out of 12) for each criterion are summarised in Table 5.17, for the All_Criteria_Met group and the equivalent group from Experiment 3.

*Table 5.17 Delayed recall performance and proportion of participants scoring at ceiling as a function of learning criterion for the All_Criteria_Met groups from this experiment and Experiment 3.*

| Condition | All_Criteria_Met group (N=52) | | Expt 3 (N=49) | |
|---|---|---|---|---|
| | *Mean(SD)* | Proportion at ceiling (12 out of 12) | *Mean(SD)* | Proportion at ceiling (12 out of 12) |
| 1 recall criterion, 55min delay | 8.23(2.55) | 11.5% | - | - |
| 1 recall criterion, 24hr delay | 7.54(2.59) | 5.8% | - | - |
| 3 recall criterion, 55min delay | - | - | 10.06(1.82) | 30.6% |
| 3 recall criterion, 24hr delay | - | - | 9.27(2.50) | 22.5% |

At both delays the 1-recall criterion keeps the proportion scoring at ceiling within both Guidelines 1 & 2, indicating the desired control of ceiling effects has been achieved. This contrasts with Expt 3, where Guideline 1 (max 50% at ceiling) is met but Guideline 2 (max 16% at ceiling) is exceeded at both delays.

The mean recall scores for each criterion and delay are also illustrated in Figure 5.7.

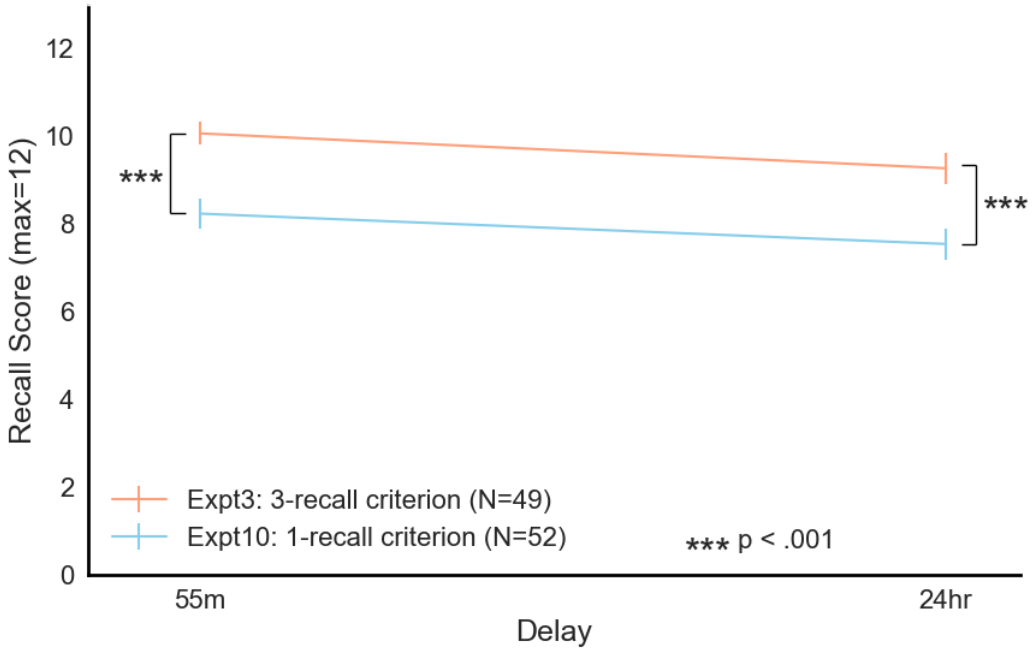


*Figure 5.7 Delayed recall performance at 55min and 24hr delays for the All_Criteria_Met group and the equivalent group from Experiment 3 (error bars +/- 1SE).*

The variation in cued recall performance between learning criteria and across delay intervals was analysed using a mixed factors ANOVA with within-subjects factor Delay (55min vs 24hr) and between-subjects factor Criterion (1-recall vs 3-recall). There were significant main effects of Criterion ($F(1, 99) = 16.08$, $p < .001$, $\eta_p^2 = 0.12$, BF = 153.4) and Delay ($F(1, 99) = 18.50$, $p < .001$, $\eta_p^2 = 0.02$, BF = 387.4) but no significant interaction ($F(1, 99) = 0.09$, $p = .77$, $\eta_p^2 < .001$, BF = 1.08). This indicates there is a significant amount of forgetting between 55min and 24hrs overall and that compared to the 3-recall criterion, the 1-recall criterion results in a statistically lower level of recall overall. However, the lack of interaction means there is no evidence of a difference in forgetting rates for different criteria.

To compare cued recall performance across criteria at each time-point two independent sample t-tests were used. The 1-recall criteria resulted in significantly lower recall at both delays (55min: $t(99) = 4.13$, Bonferroni adjusted $p < .001$, $d = 0.82$, BF = 283.6; 24hr: $t(99) = 3.41$, Bonferroni adjusted $p < .001$, $d = 0.68$, BF = 30.63), with an effect size that was large at 55min and medium at 24hr.

To further investigate the impact of criterion on possible ceiling or floor effects the distributions of scores are illustrated in Figure 5.8 for each delay and criterion, for the All_Criteria_Met group and the equivalent group from Experiment 3.

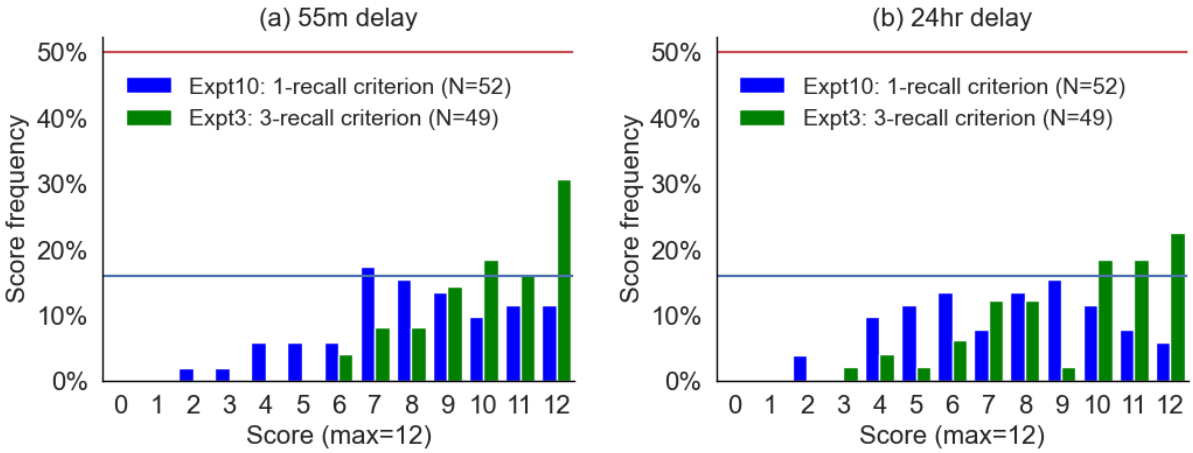


*Figure 5.8 Distribution of recall scores as function of delay and learning criterion, for the All_Criteria_Met group and equivalent group from Experiment 3. Red and blue lines represent Guidelines 1 and 2 respectively.*

At both delays the scores for the lower 1-recall learning criterion (blue bars) display a more normal distribution, and the number achieving the maximum score (12) is within both Guidelines, while the 3-recall criterion scores (green bars) fail to meet Guideline2.

The optimum result is for the scores to be as high as possible while meeting both guidelines; visually this means the proportion of participants scoring at ceiling (12) should be at, or just under, the blue line representing Guideline 2. In fact, the 1-recall scores are a little lower than optimum, especially at the longer 24hr delay. This is also reflected in the other end of the distribution, where 1 participant at 55min and 2 participants at 24hr scored 2 out of 12. These results raise the possibility that floor effects may be seen for lower performing groups such as older participants.

#### *5.5.3.3 Learning duration when learning 12 pairs to a 1-recall criterion*

If the difficulty of the VALMT learning stage has been adjusted as intended, so that most older participants can complete it within the 20min time limit, then the higher performing younger group should be able to complete it comfortably within that limit. To investigate this the total duration of learning was analysed. Figure 9 shows the distribution of learning durations for the All_Criteria_Met (Fig 9b) and larger Learnt groups (Fig 5.9a). For comparison the equivalent distributions of are also shown for the Experiment 3, which used a 3-recall criterion (Fig 5.9d & c).

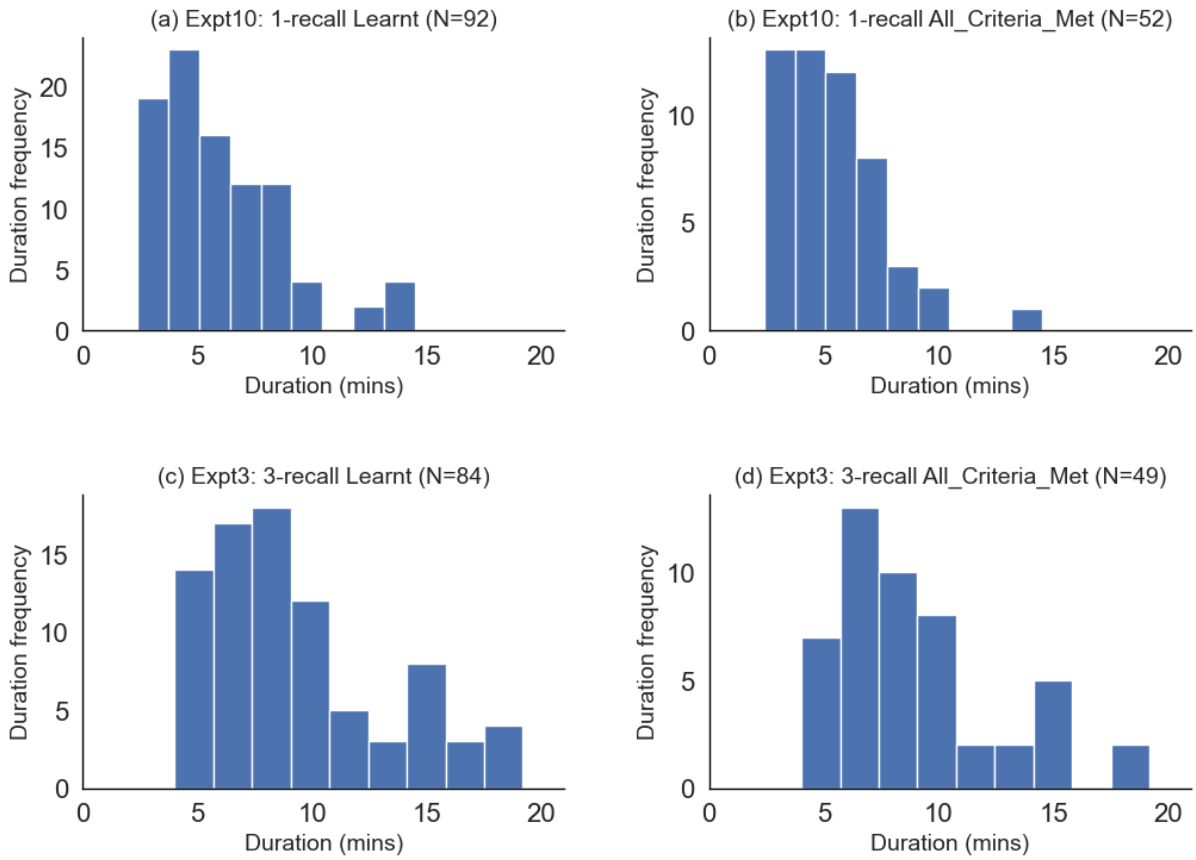


*Figure 5.9 Distribution of learning durations as function of group and learning criterion.*

The distribution of durations for both groups from the current experiment extend no further than 15min, and no participants were excluded for exceeding the 20min limit ('expired'). These results suggest older participants should be able to complete the task within 20mins, however this will need to be verified empirically in future work. By contrast, the 3-recall criterion (Experiment 3) resulted in 8 young healthy participants exceeding the 20min limit and the distributions both extend to 20mins.

#### *5.5.3.4 Recognition memory*

##### 5.5.3.4.1 Recognition memory performance

To investigate how recognition scores at 24hrs compare with the equivalent recall score the group means are shown in Table 5.18. The number of participants scoring at ceiling for each variable is also shown.

*Table 5.18 24hr recognition and recall scores for the All_Criteria_Met group.*

| Variable | All_Criteria_Met group (N=52) *Mean(SD)* | Proportion scoring at ceiling (%) |
|---|---|---|
| 24hr Recall (max=12) | 7.54(2.59) | 5.8% |
| 24hr Recognition (max=12) | 10.98(1.46) | 51.9% |

The mean recognition score at 24hr is much higher than the recall score at the same delay. This is to be expected as recognition is generally easier than recall. Over half of participants scored at ceiling for the recognition test (51.9%) indicating that the recognition test will not be able to differentiate between the top half of the young population. This will also lead to significant ceiling effects which will distort group comparisons; the proportion at ceiling fails to meet both the ceiling effect guidelines developed in Chapter 4.

#### 5.5.3.4.2 Proportion of participants scoring better on recognition than recall

One possible use of a recognition test is to monitor whether participants are paying attention and trying hard enough. If participants are paying attention they would be expected to score higher on recognition than recall, as recognition tasks are generally easier. To investigate this each participant's recall and recognition scores at 24hrs were compared. No participants scored higher on recall. This suggests all participants were paying attention.

#### 5.5.3.4.3 Items recalled but not recognised

While scoring higher on recognition than recall suggests a participant is paying attention, it is still possible that they may have failed to recognize one or more items which were recalled, as long as this was not enough to bring their total recognition score below their total recall score. To check for this the number of individual items recalled but not recognised was analysed. Across the All_Criteria_Met group (N=52) a total of 10 participants had at least one item recalled but not recognized. The maximum such items for any participant was two. The total such pairs for the whole group was 12, which is 1.9% of the total pairs tested for the whole group in the recognition test. This suggests it

is normal to see one or at most two pairs recalled but not recognised, but that it is not normal for the total recognition score to be lower than the total recall score.

#### *5.5.3.5 Recognition memory error types*

The recognition test contains two types of foils, to detect whether incorrect answers were target words from other pairs ('Intrusion'), or completely new words ('Unseen_word'). To analyse this the total errors made by the group are broken down by error type in Table 5.19.

*Table 5.19 24hr recognition error types for the All_Criteria_Met group (N=52).*

| Variable | All_Criteria_Met group errors | Proportion of total errors (%) |
|---|---|---|
| Intrusion errors | 45 | 84.9% |
| Unseen_word errors | 8 | 15.1% |

Of the three foils for each pair only one is a target word from another pair while two are unseen words, which should mean that the rate for Unseen_word errors would be twice that for Intrusion errors if participants were guessing randomly. However, the data shows most of the recognition errors are Intrusion errors, where the second word from a different pair is given as the answer; the number of Unseen_word errors is much smaller. This suggests that when participants are unsure of an answer they may responding based on familiarity, rather than guessing randomly.

#### *5.5.3.6 Memory performance of individuals with dyslexia*

The number of participants reporting dyslexia in this experiment made it possible to examine their performance as a group, compared to those from the same undergraduate population who do not report dyslexia. To investigate this, the learning and recall performance for individuals with and without dyslexia are compared in Table 5.20. Separate All_Criteria_Met and Learnt groups were prepared for those with dyslexia, where the criteria are the same as usual (All_Criteria_Met: met all criteria; Learnt: completed learning but failed to complete both delayed tests or did them at wrong time) except that the dyslexia criteria is reversed (participant must report dyslexia, rather than not report dyslexia).

*Table 5.20 Memory performance comparison for individuals with and without dyslexia, for All_Criteria_Met and Learnt groups.*

| Condition | All_Criteria_Met groups | | Learnt groups | |
|---|---|---|---|---|
| | **Without dyslexia (N=52)** | **With dyslexia (N=3)** | **Without dyslexia (N=84)** | **With dyslexia (N=7)** |
| Total Attempts Mean(SD) | 26.31(13.11) | 35.00(15.13) | 30.05(16.69) | 42.29(11.70) |
| 55min Recall Mean(SD) | 8.23(2.55) | 5.33(3.21) | - | - |
| 24hr Recall Mean(SD) | 7.54(2.59) | 4.00(1.00) | - | - |

Participants with dyslexia performed worse than those without dyslexia on both learning and delayed recall, taking more attempts to complete learning to criterion and recalling fewer pairs at both delays. For the All_Criteria_Met groups, the 55m Recall difference approaches significance ($t$ (53) = 1.89, $p = .06$, $d$ = 1.12, BF = 1.43), and for 24hr Recall it is significant before, but not after, Bonferroni adjustment ($t$ (53) = 2.34, p = .023, Bonferroni adjusted $p = .069$, $d$ = 0.1.39, BF = 2.59). For Total Attempts the difference is not significant ($Mdn_{without\_dyslexia}$ = 22.5, $Mdn_{with\_dyslexia}$ = 40.0, $MWU$ = 113, $p$ = .20, $r$ = 0.45, BF = 0.39). Although these statistics fail to reach significance, the effect sizes are all large or medium, suggesting they would be significant with larger sample sizes. Further evidence for this comes from the comparison of Total Attempts for the larger Learnt groups, where the difference is significant ($Mdn_{without\_dyslexia}$ = 26.5, $Mdn_{with\_dyslexia}$ = 45.0, $MWU$ = 456, $p$ = .016, $r$ = 0.55, BF = 0.48).

Overall, although the small sample size means conclusions drawn from inferential statistics must be tentative, these suggest that there are sufficient differences between individuals with and without dyslexia to justify excluding those with dyslexia from the main analyses.

### *5.5.3.7 Distribution of learning errors across word-pairs and their relationship to recall*

To evaluate the effectiveness of each word-pair after switching from a 3-recall to 1-recall criterion, and to investigate whether the strong relationship between learning errors and subsequent recall identified in previous experiments is retained with the lower criterion, Figure 5.10 illustrates the mean number of errors made for each individual pair, and the corresponding mean recall rate at each delay, for the All_Criteria_Met group.

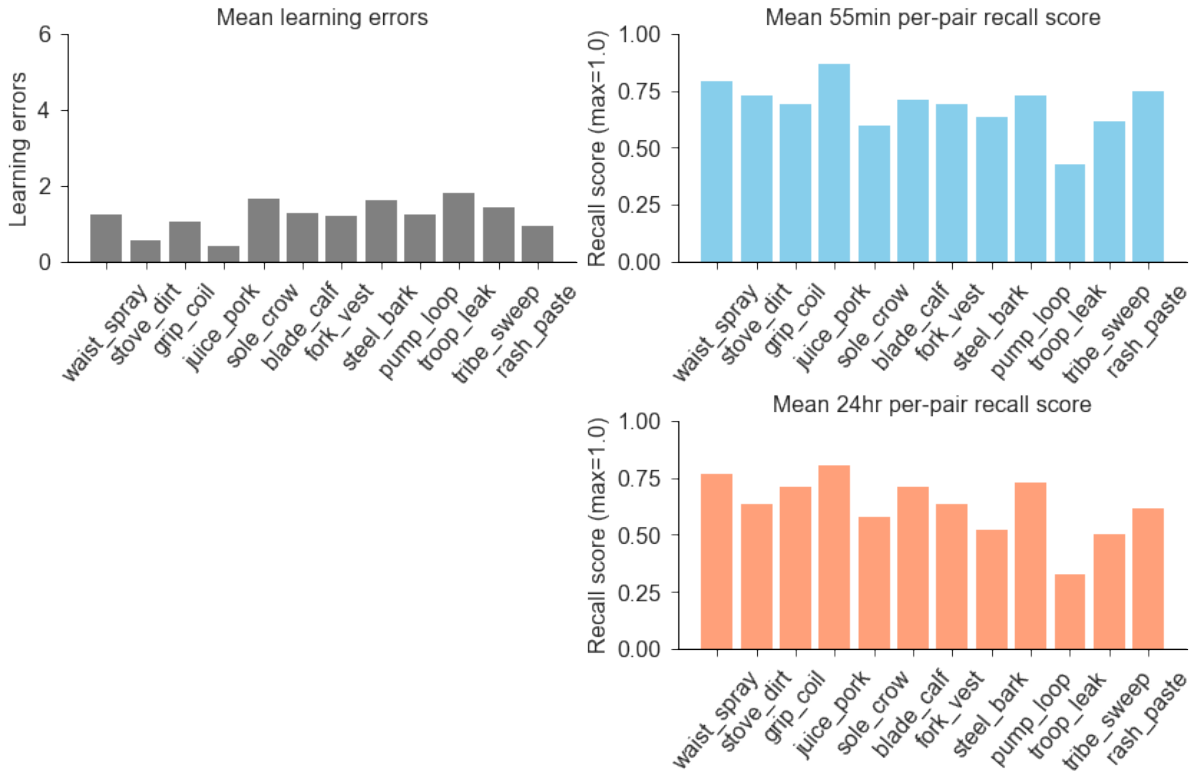


*Figure 5.10 Mean learning errors and delayed recall by individual word-pair for the All_Criteria_Met group (N=52).*

The data shows variation in difficulty between pairs which is beneficial for distinguishing between high and low performers. No pairs encounter zero or a very high number of errors, or zero or 100% recall, so all pairs are providing useful information. Visual inspection shows that in general those pairs which result in most errors are recalled more poorly at both delays, while those pairs which generate fewest errors are recalled best. In summary, the patterns observed in earlier experiments when testing with a 3-recall learning criterion are still present after changing to a 1-recall criterion

#### *5.5.3.8 Relationship between delayed recall and learning errors*

To further investigate whether the strong relationship observed in previous experiments between errors made during learning and subsequent delayed recall remains present after switching from a 3-recall to a 1-recall learning criterion, the scatterplots in Figure 5.11 illustrate the association between the total number of errors made and the cued recall scores at each delay for the All_Criteria_Met group (1-recall criterion) and, for comparison, the equivalent group from Experiment 3 (3-recall criterion).

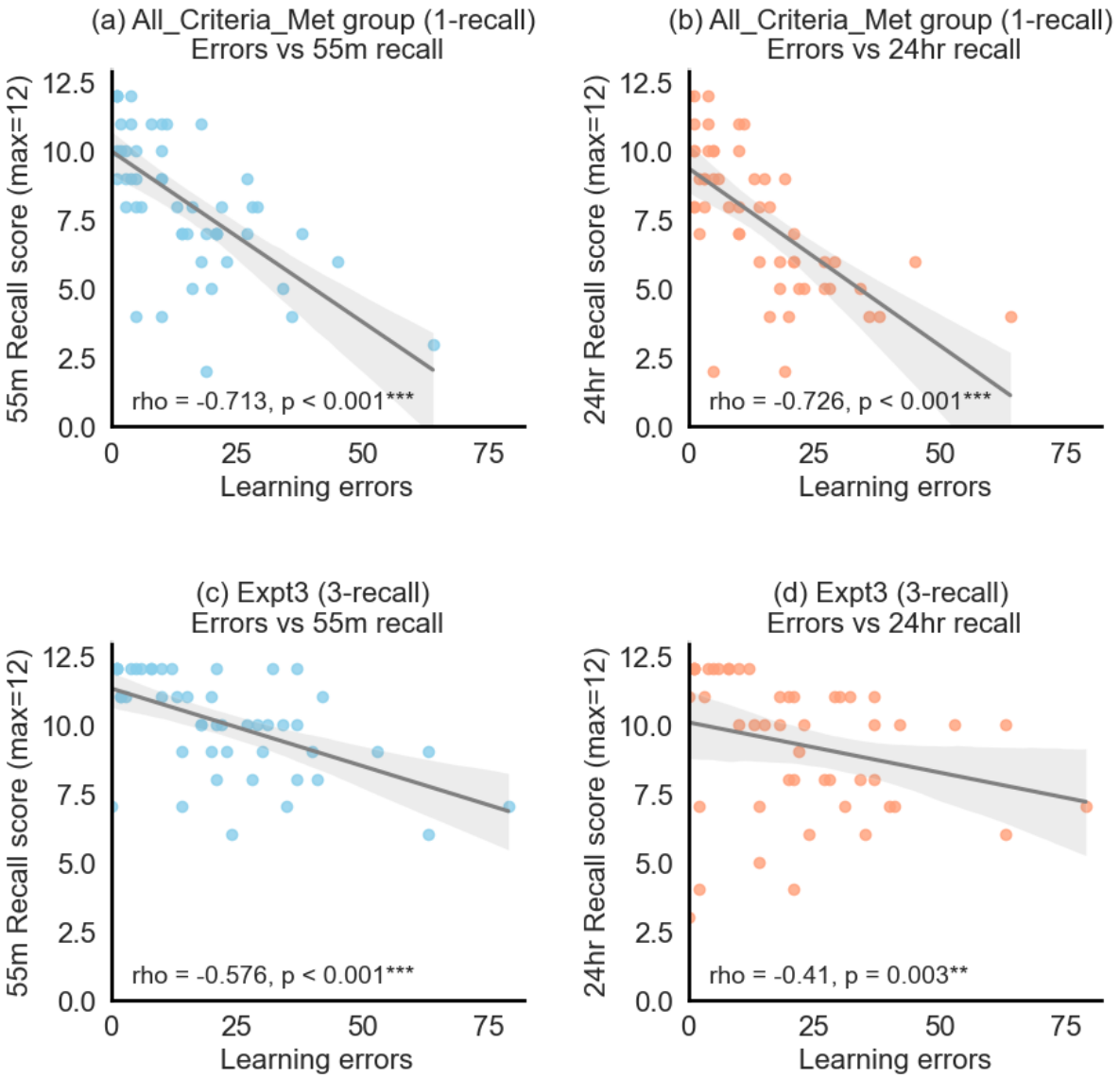


*Figure 5.11 Correlation of total learning errors with delayed recall for the All_Criteria_Met group (N=52) and equivalent group from Experiment 3 (N=49); shaded area is 95% confidence interval.*

There was a significant negative correlation between learning errors and delayed recall at both 55min and 24hr (55min rho = -0.71, $p < .001$; 24hr rho = -0.73, $p < .001$), with those who make the most errors subsequently recalling the fewest pairs. This indicates that the strong relationship between these variables observed in Experiment 3 is still present after changing to a 1-recall learning criterion. In fact, the relationship has become more obvious in the scatterplots, and the correlation coefficients are larger, due to the greater overall variation in recall scores once ceiling effects are minimised. This reveals another example of how ceiling effects in earlier experiments may have been hiding relationships or causing them to be underestimated.

To investigate this at a per-pair granularity the scatterplots in Figure 5.12 illustrate the relationship between the mean number of errors made when learning each individual pair

and the subsequent mean recall scores for that specific pair. These plots show one data-point for each of the 12 word-pairs. Results are plotted from the All_Criteria_Met group (1-recall criterion) and the equivalent group from Experiment 3 (3-recall criterion).

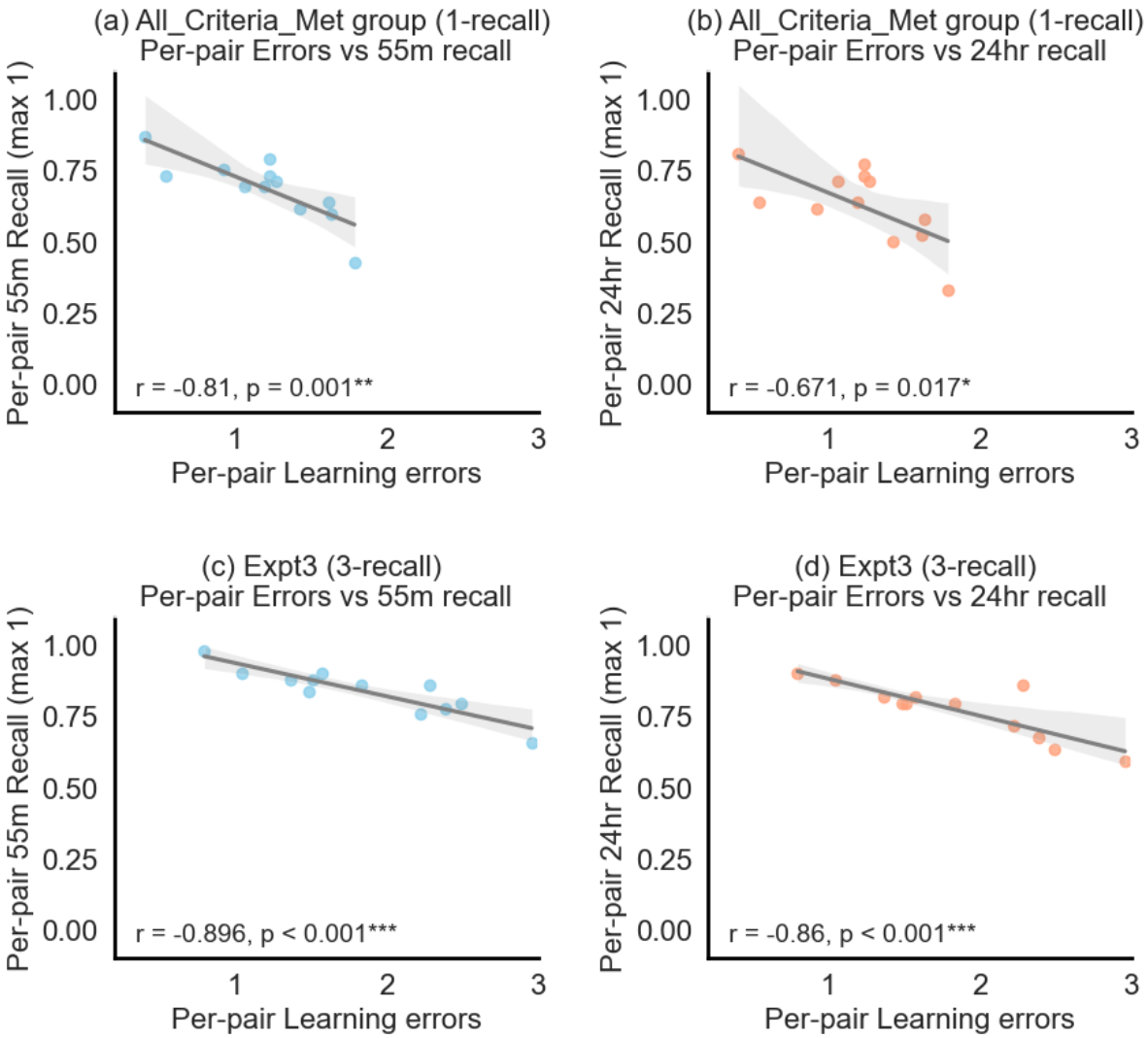


*Figure 5.12 Correlation of learning errors per word-pair with delayed recall per word-pair for the All_Criteria_Met group (N=52) and equivalent group from Experiment 3 (N=49); shaded area is 95% confidence interval.*

The large and significant negative correlation between the per-word pair learning errors and delayed recall that was observed in Experiment 3 is still present at both 55min and 24hr (55min r = -0.81, p = .001, BF = 32.6; 24hr r = -0.67, p < .001, BF = 4.58). This confirms the visual pattern seen in Figure 5.10, in which those pairs which encounter most errors are recalled most poorly.

### 5.5.4 Discussion

#### *5.5.4.1 Optimising learning criteria to reduce risk of ceiling effects*

The previous experiments in this chapter showed that it would be necessary to retain the repeated recall for the delayed tests, as it was not feasible for lower performing participants to learn sufficient word-pairs to provide enough unique pairs to be recalled at each delay. They also indicated that to minimise ceiling effects for the younger participants at the shorter delay the optimum learning criteria was 1-recall. This experiment is the first to combine these parameters, with 12 pairs being learnt to a 1-recall criterion and then recalled at both delays.

The optimum result would be for the delayed recalls scores for the young healthy participants in this experiment to comply with both guidelines at both delays, but not to be any lower, as this would provide the maximum range of lower scores available for lower performing groups before they are impacted by floor effects. In practice, the delayed recall scores at both delays did comply with both ceiling effect guidelines, indicating that ceiling effects had been controlled. However, the scores were a little lower than optimum, especially at 24hrs. This indicates that there is a risk that scores for lower performing groups such as older participants may suffer from floor effects. Future work should include testing with lower performing groups to see if floor effects do in fact occur. If the 1-recall criterion does result in significant floor effects for lower performing groups then the next option would be to switch to 2-recall criteria, or perhaps consider a mixed criteria design in which half the pairs are learnt to a 1-recall criterion and half to a 2-recall criterion but all pairs recalled at both delays.

In summary, current evidence suggests a 1-recall criterion is optimum for a general purpose VALMT to be used with mixed ability levels; however monitoring for floor effects with lower performing groups should continue.

#### *5.5.4.2 Learning duration when learning 12 pairs*

Since the experiments in the previous chapter showed that 16 pairs was too many for older participants to learn within 20mins, this experiment reverted to using a set of 12 word-pairs. However, this is the first time all twelve have been learnt using a single 1-recall learning criterion. One aim of the current experiment was to verify that these could be learnt by these young participants within the time limit. In fact, to be sure that lower

performing groups such as older participant can complete learning, younger participants should be comfortably within the time limit, as lower performing groups will take longer.

In practice, all young participants completed learning in under 15mins, suggesting it should be possible for the vast majority of older healthy participant to complete learning within the 20min limit. However, this needs to be verified empirically in future work. If results show that even older participants can complete leaning of 12 pairs to a 1-recall criterion comfortably within 20mins then it may be possible to move to 14 word-pairs, which would improve sensitivity and reliability, or alternatively reduce the time limit from 20mins to a shorter interval. Using a shorter time limit would be beneficial as it would reduce the total time the test takes to administer in a clinical environment, and it also likely to cause fewer participants to dropout in general online testing through tedium or fatigue.

#### *5.5.4.3 Forgetting rate between 55min and 24hr*

The previous experiments in this chapter investigated the option of removing the repeated recall for the 24hr delayed test, with the aim of revealing more forgetting over the 55min to 24hr delay. This option proved impractical. For this experiment the repeated recall was retained, but the learning criterion was reduced to 1-recall to lower scores. This was done primarily to control ceiling effects, however, a second potential benefit was that reducing the strength of memories might result in greater forgetting, providing an alternate way to expose forgetting between 55min and 24hr.

However, the results showed little difference between forgetting rates with a 3-recall and a 1-recall criterion (Experiment 3 and Experiment 10 respectively), with the forgetting rates for both criteria being low. While this means the observed forgetting rate for young healthy participants remains low, it does align with existing published work; both Slamecka and McElree (1983) and follow-up work by Rivera-Lares et al. (2022) found that forgetting rates are not impacted by varying the initial degree of learning, although it should be noted that their work only tested with young participants (18-30yrs). Future work should investigate whether the change in criterion has a greater impact on 55min to 24hr forgetting rates for older participants.

#### *5.5.4.4 Recognition memory*

This was the first experiment to use a version of VALMT that included a recognition memory test. This four option forced-choice recognition test (4FCR) was completed after the final 24hr delayed recall test, to avoid disrupting any of the other elements of the overall test.

Group mean recognition scores were higher than recall scores. In fact, over half of participants scored at ceiling (12 out of 12) for recognition. This means for high performing groups such as young healthy participants there will be significant ceiling effects which may distort group comparisons. At an individual level the fact that so many higher performers score at ceiling means it will not be possible to differentiate between recognition performance for many participants. However, as the VALMT is designed to detect subtle deficits in memory rather than differentiate between the highest performers, this does not invalidate the recognition test. In addition, even when a participant scores at ceiling, this still shows that their recognition memory is better than their recall memory performance, and that they are paying attention and exerting adequate effort. That participants had been paying attention was further confirmed by comparison of 24hr recall and recognition scores at an individual participant level, which showed that no participants scored lower on recognition than on recall. However, at the individual item level a very small number of items were recalled but not recognised (1.9%). The maximum of such items for any participant was two. This may indicate a lapse in concentration, or could reflect competition between the correct answer and the foil which is a target word from another pair. Either way, the data suggest that two criteria to identify young healthy participants who are not paying attention would be scoring more on recall than recognition, and failing to recognise more than two pairs that had been recalled. These same criteria could potentially be used to identify malingering, since malingerers are likely to deliberately choose foils rather than correct answers. Assessing whether these same two criteria are valid for older participants will require further testing.

The types of errors made during the recognition test indicated that when participants made an error they generally fell back on familiarity. They were much more likely to select the foil which was a target word from another pair (Intrusion error), so had been previously seen, than either of the two foils which were previously unseen words (Unseen_word error). In contrast, if they were guessing randomly the frequency of Unseen_word errors would have been twice that for Intrusion errors.

One older participant who had been impacted by encephalitis was excluded from the main analyses due to only managing to learn 6 pairs to criterion. They were subsequently unable to recall any pairs at either delay, yet they scored 11 out of 12 on recognition. This single case evidence provides an example of how the recognition test may be useful in analysing the nature of deficits displayed by more impaired individuals or groups.

#### *5.5.4.5 Performance of individuals with dyslexia*

The number of participants reporting dyslexia in this experiment made it possible to examine their performance as a group for the first time. Although the sample size was small, the data suggests that individuals with dyslexia display lower memory performance, requiring more attempts to complete learning, and recalling fewer items at delayed test. Since the sample size is small, and the demographics are limited to a simple Yes/No choice (no data on severity was collected), it is not possible to examine this more closely as part of this experiment. However, the data are sufficient to indicate that individuals with dyslexia should be excluded from the main analyses, as has been done in all experiments so far. This also has clinical implications, meaning that VALMT should not be used with individuals with dyslexia without preparing a separate set of norms. This is one inevitable characteristic of a test that requires participants to read the material being learnt, rather than having it read aloud to them.

#### *5.5.4.6 Distribution of learning errors across word-pairs and their relationship to recall*

Lowering the learning criterion from 3-recall to 1-recall for all pairs did not change the effectiveness of the selected pairs, or change the strong relationship between learning errors and subsequent recall. The results still showed a similar variation in difficulty between pairs which is beneficial for distinguishing between high and low performers, and no pairs encountered zero or a very high number of errors, or zero or 100% recall, so all pairs continue to provide useful information. Those pairs which resulted in most errors were still recalled more poorly at both delays, while those pairs which generate fewest errors were recalled best.

#### *5.5.4.7 Relationship between delayed recall and learning errors*

The strong relationship between errors made during learning and subsequent delayed recall which was observed in previous experiments remained after changing the learning

criterion to 1-recall. In fact, the observed relationship between total learning errors and delayed recall was even stronger. This was due to the elimination of ceiling effects. In earlier experiments many participants who made a low number of errors scored at ceiling, which meant variation in errors for these higher performing participants was unable to further influence their recall score, reducing the apparent correlation. After minimizing ceiling effects by changing the learning criterion there was greater overall variation in recall scores and the observed correlation coefficients were larger. This reveals another example of how ceiling effects in earlier experiments may have hidden relationships, or causing them to be underestimated. Overall, after lowering the learning criterion the errors made during learning continued to provide a strong indicator of subsequent delayed recall.

### 5.5.5 Conclusion

This experiment found that a 1-recall learning criterion keeps ceiling effects within guidelines for young healthy participants when tested at 55min and 24hr, with the same material tested at both delays. However, especially at 24hr, the scores were a little lower than optimum, which may indicate a risk of floor effects with lower performing groups. This should be monitored in future when testing with older age groups. The lower learning criterion did not result in greater forgetting between 55min and 24hr for this age group. Future work should investigate whether the impact on forgetting between these intervals is greater for older participants. The previously identified strong relationship between errors during learning and subsequent recall remained after lowering the learning criterion; in fact the reduction in ceiling effects made this relationship more obvious.

The new VALMT recognition test worked successfully. Although mean recognition scores were high, the test is a valid way to assess recognition memory for lower performers, and criteria were identified to enable it to be used as a marker for lack of attention and malingering.

Finally, individuals reporting dyslexia displayed lower memory performance, confirming it is valid to use dyslexia as an exclusion criteria in general VALMT research.

## 5.6 General Discussion

### 5.6.1 Impact of removing the repeated recall

Experiments 3 to 6, the first to use the new online VALMT, found that although older participants forgot more between 55min and 24hr, the forgetting rates were small for all age groups. One reason for this could be that the same word-pairs were recalled at both delays, with the recall at 55min providing retrieval practice and strengthening memories. Experiments 7 to 9 in this chapter investigated this by removing the repeated recall, with different pairs recalled at each interval. They found evidence that non-repeated recall at 24hr was lower than the equivalent repeated recall at the same delay, for both age groups. This confirms that recalling the same material at both delays does lead to greater retention. While this result provides useful information on how VALMT could be tuned to increase sensitivity, it also provides further evidence that recalling the same material at each delay, as used in many published research studies, may hide forgetting that would otherwise be visible, or lead this to be underestimated.

### 5.6.2 Feasibility of learning 16 pairs to reduce need for repeated recall

Although it appears that recalling different material at each delay would provide a more sensitive test, this will only be feasible in practice if it possible for participants to learn enough pairs to provide unique sets for each delay. Experiments 7 to 9 investigated whether it would be possible to increase the number of pairs learnt during the learning stage of the online VALMT from 12 to 16, which would provide 8 pairs to recall at each delay. While younger participants were able to learn this many pairs within 20min, it was on the limit of what lower performing younger participants could do. Sixteen pairs proved too many for some of the older participants, and many others came close to the time limit. Considering that these older participants were healthy, it can be expected that those with any memory impairment would find this task even harder. This suggests that if a single version of the VALMT is to be used with multiple groups with different abilities then the number of pairs will need to be limited to the original 12. If different pairs were to be recalled at each delay this would only provide 6 pairs for each delay, which would negatively impact test sensitivity and reliability. As a result, it will not be feasible to remove the repeated recall, and other approaches should be investigated to increase test sensitivity.

### 5.6.3 Optimising learning criteria to reduce risk of ceiling effects

The analysis in the previous chapter highlighted that some VALMT experiments may have been impacted by ceiling effects, and that steps should be taken to reduce these. The experiments in this chapter investigated whether reducing the learning criterion could achieve this, making the test harder and ensuring that fewer participants achieve the maximum score at delayed recall. They showed that this approach was effective, and that the optimal criteria for material to be recalled for the first time at 30min and 24hr were 1-recall and 3-recall respectively. These values ensured that both ceiling and floor effects were avoided for healthy older participants. However, for younger participants scores complied with Guideline 1, but exceeded Guideline 2. This suggests there may still be some impact of ceiling effects for high performing groups even with these criteria.

The difference between the Younger and Older groups' delayed recall at 30min also provided the first empirical evidence that reducing ceiling effects may make the test more sensitive, as intended. However, although the group difference was larger than in Experiment 4 which used a 3-recall criterion, it was not large enough to be statistically significant, indicating that the first delayed test should continue to be performed at the longer 55min delay rather than 30min, to provide more time for forgetting to manifest.

These experiments also identified an interesting variation in how different age groups were impacted by varying the learning criterion. A higher learning criterion led to an increase in the total errors made during learning. However, this impact was larger for older participants. This suggests that younger participants can benefit from the retrieval practice arising from a higher learning criterion with little negative impact from increased learning errors, while for older participants the greater increase in learning errors may reduce the benefit they get from the additional retrieval practice. This difference in impact across age groups reflects the lower learning performance of the older participant group, and a similar outcome would be expected for any group or individual with a learning deficit. If this analysis is correct then a higher learning criterion may actually, in some situations, accentuate observed group differences in delayed recall.

These optimised learning criteria were selected for a design which eliminated the repeated recall at 24hrs. Since these experiments have shown that it will be necessary to continue with a repeated recall, this raises the question of what criterion should be used for such a design. In a repeated recall design all pairs are learnt to the same criterion and are recalled at both delays. To achieve the best control of ceiling effects for the shorter

delay, the 1-recall criterion should still be optimal, as this is the first delayed recall. If this criterion is chosen, what performance would be expected for the repeated recall at 24hrs? Compared to the conditions used for 24hr recall in experiments 8 and 9 in this chapter (3-recall criterion, non-repeated recall) the change to a 1-recall criterion should reduce scores, while the repeated recall should raise them. The resulting incidence of ceiling and floor effects was evaluated empirically as part of validating an optimised VALMT in Experiment 10.

### 5.6.4 Optimised VALMT design

Based on the results of experiments in this and previous chapters, an optimised VALMT for general purpose use was created and tested in Experiment 10, using a single set of 12 word-pairs, learnt to a 1-recall criterion, with all pairs recalled at 55min and at 24hrs. The resulting delayed recalls scores at both delays complied with both ceiling effect guidelines, indicating that ceiling effects had been successfully controlled. Scores were a little lower than optimum, suggesting there may still be some risk of floor effects for lower performing groups; future work should monitor this closely. The previously identified relationships between errors made during learning and subsequent recall were still present after switching to the optimised design, so the previous analyses involving the role of errors will still be valid.

Importantly, although reducing the learning criterion from 3-recall to 1-recall lowered delayed recall scores at both delays, it made no difference to the forgetting rate between 55min and 24hr. This result is in agreement with early work by Slamecka & McElree (1983) and more recent follow-up work by Rivera-Lares et al. (2021) that found changing the initial learning level for verbal material does not alter the subsequent forgetting curve.

### 5.6.5 Recognition memory test

A recognition test, added after the second delayed recall, worked successfully. For young healthy participants, a significant number achieved the maximum score, so there will be ceiling effects for high performing groups. However, even for higher performers the test will still provide a valid marker for attention and effort.

Analyses of error types indicates that when young healthy participants make an error they are more likely to choose a previously seen, but incorrect, word rather than a previously unseen word. This suggests that when unsure of the correct answer they rely

on familiarity to guide their choice. Future testing with older participants or patient groups could investigate whether they display the same pattern. There may also be value in adding a Remember/Know/Guess check (RCA paradigm; Gardiner & Java, 1993) for each recognition answer, allowing changes in the subjective nature of recognition memory to be investigated, even where this has not reached the point of leading to a recognition failure.

# 6 Chapter 6 – General discussion

This thesis has detailed the development of a novel memory test, the VALMT, and its use in addressing some open questions in the area of accelerated forgetting. An initial study space analysis identified the need for additional ALF research beyond the original focus on epilepsy, especially investigating ALF prevalence and symptoms in healthy ageing (above 60yrs) and the possibility of using ALF as a marker for those at risk of developing MCI/AD. There was also a need to develop an ALF measure which avoids the methodological flaws in existing work, and could be used in both research and clinical applications.

Initial experiments using the original VALMT (Experiments 1 and 2) showed that it could reveal differences in delayed recall performance between younger and healthy older participants at an earlier time point than was possible using a standard clinical test. Importantly, older participants who learnt more slowly also forgot more rapidly within the window of a standard clinical visit, and this objective impairment was associated with increased subjective memory complaints, indicating a possible link to pre-clinical Alzheimer's disease or other dementia processes. However the procedure used was complex to administer, limiting its usefulness in clinical settings, and introducing possible confounds due to interference.

A fully online version of VALMT using a simplified procedure overcame many of the practical limitations of the original face to face testing procedure, allowing wider scale testing without the need for travel. The simplified procedure also eliminated the complex interleaving of learning and testing, removing this as a possible confound. A number of experiments showed this new version was effective for all age groups (Experiments 3 to 6). These experiments replicated the same general pattern seen with the original face-to-face version (Experiments 1 & 2), although the effect sizes seen were smaller. The biggest predictor of accelerated forgetting was the number of errors made during learning. Importantly, once this factor was equalised there was no evidence of a greater underlying forgetting rate in slow learning older participants, suggesting their main problem may be a learning deficit rather than retention. In this model, some older participants have a cognitive deficit which causes them to make more errors during learning to criterion, and these errors cause interference leading to greater forgetting. These experiments also indicated that in the general population, memory performance starts to decline around the mid-fifties.

Results from the first VALMT experiments also showed evidence of ceiling effects for higher performing groups. Due to a lack of practical guidelines in the existing literature, an analysis using a computer modelling approach was performed to investigate the impact of ceiling effects on the results of experimental designs typical of ALF research. As an outcome of this a set of practical design guidelines were developed, and the online VALMT was modified accordingly. A set of experiments validated an optimised VALMT design, which brought ceiling effects within the guidelines.

These experimental findings and their research and clinical implications are discussed in the following sections, in relation to the research topics identified for this PhD.

## 6.1 Development of VALMT as a measure of ALF

A number of improvements were made to the VALMT over the course of the work detailed in this thesis, aimed at improving sensitivity and making it more practical for general purpose research and clinical use.

### 6.1.1 Online operation

One major change was the development of an online version. The fact that this version worked successfully across a wide range of ages suggests it is viable as a general purpose testing tool. This has implications for future research, with many benefits. The process is fully automated, which standardises the testing procedure and eliminates risk of experimenter bias. It also removes the need for careful training of researchers and clinicians, making it simpler and faster to run studies, or apply the test in a new clinical setting. This also improves scalability, as the number of participants in a study is no longer limited by the number of researchers available. The use of an internet based solution supported on any device with a web browser makes the test accessible to the majority of the general population, without the need for any travel on the part of the participants, or researchers and clinicians. This greatly expands the opportunities for application of the test, and should facilitate easier recruitment of participants. It also increases the feasibility of repeated testing and testing at longer delays. A final benefit of online operation became evident during the Covid-19 pandemic; remote testing in this manner can continue regardless of restrictions on travel and face-to-face contact.

However, the online version did suffer from a higher attrition rate than face to face testing, with some people dropping out before completing all stages, or completing a

delayed test outside the normal acceptable window. This is perhaps to be expected, as it is easier for participants to drop out when they have no direct face-to-face contact with a researcher, and feel less obligation to continue. This highlights the need for some form of automated reminder, which will send an email or text message to participants encouraging them to complete their delayed tests and to do this at the correct time.

### 6.1.2 Ceiling effects and learning criteria

The computer modelling approach used in the current research made it possible to simulate thousands of experiments, and evaluate the impact of various levels of ceiling effect. This showed that such effects can cause issues for ALF research in two ways. First, they can reduce the observed difference in group means when comparing higher and lower performing groups at short delays, potentially lead to the failure to detect a difference where one exists (false negative). Second, they can increase the observed group differences in forgetting rates between a short and long delay, potentially leading to detection of a difference where none exists (false positive). It also allowed practical guidelines to be developed for the design of ALF experiments, to ensure ceiling effects remain with tolerable limits, where they should not significantly impact results.

A review of a seminal study in the ALF literature (Butler et al., 2007) confirmed that ceiling effects are likely to have influenced results, leading to an underestimation of the difference in performance levels between controls and a patient group (TEA patients) at a 30min delay, and an overestimation of the difference in forgetting rates between 30min and one week. While in this specific study ceiling effects seem unlikely to have led to a false claim of an accelerated forgetting rate or changed the result of any statistical tests of significance, it nevertheless may have impacted the strength of the authors conclusions. An important part of the argument used in early work for the existence of ALF is that certain groups show normal learning and retention to a short delay of at least 30mins, followed by accelerated forgetting at longer delays. If the difference in group performance at the short delay has been underestimated, then this suggests that forgetting may start earlier than suggested by the proponents of ALF as a late-onset forgetting (Butler et al., 2007; Mayes et al., 2019), strengthening the arguments of those such as Cassel and Kopelman (2019) who argue forgetting actually starts early and that this can be detected when methodological weaknesses are eliminated. The guidelines developed in this thesis could be used by other researchers to review the possible impact of ceiling

effects on their historical results and to keep ceiling effects within tolerable limits in future studies. This would be of particular benefit to attempts to theorise the timeframe of onset for ALF, and its existence as a qualitatively unique form of forgetting.

The retrospective analysis of the Butler at al. (2007) results was constrained by the data available in the published paper; only summary statistics (means and standard deviations) are available for each group. In light of this, a further recommendation would be that authors should publish the percentage of participants who score at ceiling (or floor) at each delay. This would allow the level of ceiling effects to be evaluated against the guidelines.

Applying the guidelines retrospectively to previous VALMT experiments showed that the level of ceiling effects was high enough in some cases that it may have influenced results. In an effort to prevent this happening in future, changes were made to the online VALMT to reduce the mean scores, and thereby reduce the proportion of higher performing groups that achieve the maximum score. Experiments 7 to 10 showed it was possible to reduce scores by changing the learning criterion. Early VALMT studies used a 3-recall criterion, requiring each word-pair to be correctly answered three times during the learning phase. Changing this to require two recalls or one recall was shown to reduce scores. To control ceiling effects for a short delay of 30min or 55min the optimum criterion was shown to be 1-recall. This kept the scores for young healthy participants within, or very close to, the guidelines. Since this is expected to be the highest performing group this means other groups can expected to score lower and therefore also meet the guidelines. As a result this version of VALMT is well suited to testing with a wide range of ages and populations in future.

Although adjusting the learning criterion reduced scores, it did not change the subsequent forgetting rate between 55min 24hrs. The question of how the initial level of learning influences subsequent forgetting has been relatively ignored in memory research. However, some initial work by Slamecka & McElree (1983) and more recent follow-up work by Rivera-Lares et al. (2021) had suggested that changing the initial learning level does not alter the subsequent forgetting curve. These new results from adjusting the learning criterion of VALMT are in agreement with this previous research. This has implications for other researchers as it provides further evidence that equating initial learning through learning to a criterion does not impact the subsequent forgetting curve.

### 6.1.3 Eliminating interleaving of learning and testing as a potential source of interference

One possible explanation for the significant differences found between younger and older groups and between slow and fast learning older participants in Experiments 1 and 2 is that the procedure involved a relatively complex interleaving of learning and testing. This may have created interference, and if some older participants were more vulnerable to interference this could explain their subsequent faster forgetting. One aim of the simplified procedure in the online VALMT was to remove this possible source of interference by using a simplified procedure with a single learning phase.

Comparison of results from the original version and the new online VALMT found a smaller effect size for between-group comparisons with the online version, suggesting interference may indeed have been a factor in the early experiments. More detailed analysis suggested that while this interference factor did not impact the Younger age-group, it may have impacted the Older age-group, although it is not clear to what extent the reduced group differences in the online version were caused by the removal of such interference, or by a difference in the learning performance of the Older samples across studies. This has implications for future research; although further work will be required to strengthen this conclusion, it seems that future experiments should avoid the interleaving of learning and testing.

### 6.1.4 Use of repeated recall for the 24hr test

For the online VALMT there is just one learning phase, with all pairs learnt together. Testing showed that while 12 pairs could be learnt by both younger and older participants within 20mins, 16 pairs was too many for older participants to learn within that time limit. This means that to have a version of VALMT which can be used with all ages it is necessary to limit the test to 12 pairs.

While it would be preferable to test different pairs at each delay, to avoid the retrieval practice that comes from repeated recall of the same pairs, this is only feasible if sufficient pairs can be learnt to provide a unique set per delay. With 12 pairs and two test delays only six pairs are available per delay, which will reduce test sensitivity and reliability. For this reason the decision was taken to use a repeated recall for the second delay for the online VALMT, so that all 12 pairs are recalled at both delays. It was hoped that testing at 24hrs might provide enough time for forgetting to become detectable, despite any retrieval practice, in those with a long-term forgetting deficit.

While this repeated recall is sub-optimal, it reflects a common compromise made with this type of testing. In a recent attempt to overcome this problem Baddeley et al. (2019) developed their Crimes test. This involves memorising a series of short events, with 4 different questions being possible for each event, by probing a different aspect. This was intended to provide material which could be comfortably learnt within a single learning phase, but provide sufficient unique test questions for four test delays. However, they found that probing one aspect of an event primed the other elements, reducing the observed forgetting. At this time there is still no accepted test that allow testing of verbal material at multiple delays with unique stimuli for each delay.

### 6.1.5 Forced choice recognition testing

The literature review of Chapter 1 highlighted that it would be beneficial for ALF testing to measure both recall and recognition memory performance. This would allow researchers and clinicians to look for any differential impact of ALF on recollection and recognition processes. For this reason a forced choice recognition test was added to the online VALMT, operating after the second delayed recall test to avoid any possible impact on recall testing.

All participants scored higher on recognition than recall. This was predicted, as recognition is generally easier than recall. However, this also confirms that scoring lower on recognition than recall could be used a marker for a lack of attention/effort, or active malingering. This will be particularly useful for online testing where it is not possible to judge attention levels by observing the participant.

Where participants made errors in recognition they generally fell back on familiarity, choosing a word they had previously seen in another pair over a word they had not previously seen. This supports a dual process model of recognition, with recognition having both a recollective and familiarity component (Diana et al., 2006; Yonelinas, 1994).

The high scores achieved for recognition meant there was no evidence of accelerated forgetting impacting recognition memory. However, testing was only performed with younger participants; it remains possible that future testing with older participants may find such evidence. However, to allow more detailed analysis of recognition memory despite high scores there may be value in adding a Remember/Know/Guess check (RCA paradigm; Gardiner & Java, 1993) for each recognition answer. This would allow any

changes in the nature of recognition memory to be detected even where this has not reached the point of leading to a recognition failure.

Although not included in the main analyses of the reported experiments, a single older participant who had been impacted by encephalitis was observed to be very impaired on learning and recall, managing to learn only 6 pairs to criterion and scoring zero on delayed recall at both delays, yet they scored 11 out of 12 on recognition. This participants poor learning and recall performance combined with good recognition performance illustrates how the recognition test may be useful in analysing the nature of deficits displayed by more impaired individuals or groups.

## 6.2 ALF in healthy ageing, and possible use of VALMT as a marker for risk of MCI/AD.

A common pattern of results was observed across multiple experiments using the two different versions of VALMT, with both face-to-face and online VALMT versions able to identify differences in delayed recall performance between younger and older participants, in some cases within 55min which is sooner than a standardised clinical measure (WMS-LM) could. This between group difference in recall was driven by slower learning older participants, who performed more poorly than their fast learning peers, and being a slow learner was associated with increased memory complaints.

However, although the general pattern was the same across versions, the effect sizes observed and the delay at which differences became statistically significant were different. Initial testing using a face-to-face version of VALMT and testing at 5min, 30min and 55min found a large difference in recall between younger and older participants by 55mins, and evidence of ALF between 30min and 55min for the slow learning older participants. Subsequent testing using the new online VALMT and testing at 55min and 24hr found that older participants again scored lower on delayed recall by 55min, but the difference did not reach significance until 24hrs. There was also weak evidence of ALF between 55min and 24hr in the slow learning older participants, however this only approached significance ($p = .07$).

The smaller effect size seen at 55mins was due to the older group performing better in the online experiment. A range of possible explanations for this were analysed. This analysis concluded that the most likely causes were the exclusion of the worst performers in the online experiment due to a 20min limit imposed on learning to criterion, and the removal of interference due to the simplified procedure which eliminated the interleaving

of learning and testing used in the face-to-face testing. Subsequent changes made to the VALMT during tuning to avoid ceiling effects are expected to increase the effect size observed at 55min in future online studies in two ways. First, eliminating ceiling effects for the higher performing group (younger participants) should avoid suppression of the mean score for this group, resulting in a larger observed difference in group means. Second, switching to a 1-recall learning criterion will reduce the total time taken to complete learning, resulting in fewer of the lowest performing older participants being excluded because they fail to complete learning. It is predicted that these changes will enable the online version to detect statistically significant differences between younger and older participants within 55min.

With regard to ALF, the face-to-face experiments found clear evidence of accelerated forgetting in the slow learning older participants with a relatively early onset, between 30min and 55min. For the online experiments, the first test delay was 55min, so any ALF starting within the same timeframe will already have started by the first delayed test. This complicates the interpretation of any differences in forgetting rates between 55min and 24hr; a faster forgetting rate may just reflect a continuation of a faster forgetting which started between 30min and 55min, rather than reflecting ALF with an onset between 55min and 24hr. When only those older participants who perform normally at 55min (within 1SD of the group mean) are included any evidence of late onset ALF disappears. Taken together the results across multiple experiments indicate that slow learning older participants display ALF with an early onset, prior to 55mins, but do not display any additional late-onset ALF (beyond 55mins). The majority of previous work investigating ALF in ageing or MCI/AD has tested at 30mins, followed by one extended delay of several days or even weeks, and found evidence of ALF between these two delays. The current results are compatible with existing published work, but suggest that if researchers had tested at shorter extended delays they may have detected ALF within a shorter timeframe.

The fact that slow learning older participants forget more rapidly makes it important to understand why some older participants learn more slowly and therefore make more errors during learning. Analysis ruled out age, attention, ability to understand instructions, and familiarity with computers (time taken per attempt used as a proxy) as primary causes of learning errors, leaving some form of cognitive deficit as the most likely explanation. Multiple indirect lines of evidence suggest such cognitive decline may be a pre-clinical stage of AD or another cause of dementia. First, there is a correlation

between VALMT learning errors and subjective memory complaints, a known marker for risk of MCI/AD (Mitchell et al., 2014; Weston et al., 2018). Second, the type of associative memory tested by VALMT relies on hippocampal and entorhinal cortex regions which are the first to be damaged by the earliest stages of AD (Sapkota et al., 2017; de Rover et al., 2011). Third, episodic memory is one of the first cognitive functions impacted by AD (Fox et al., 1998), and impairment of this sort is one of the criteria used to diagnose amnestic MCI (aMCI) which carries an increased risk of progression to AD (Silva et al., 2012). In summary, several indirect lines of evidence suggest the possibility that VALMT learning performance may provide a marker for those at risk of developing dementia. More direct evidence could be found in future by comparing MCI patients performance on VALMT and the existing clinical tests which are one of the factors considered when diagnosing this condition, to see if they concur, and through longitudinal experiments to evaluate the ability of VALMT to predict future progression to MCI or AD.

The fact that older participants who make more errors during learning subsequently forget more rapidly also makes it important to understand the relationship between learning errors and recall; specifically, do errors lead to poorer recall, or does a third factor such as a faster underlying forgetting rate cause both? Across multiple experiments the strongest predictor of delayed recall performance was the number of errors made during the learning process, with those who made the most errors subsequently recalling the fewest pairs, and the pairs which generated the most errors being recalled most poorly. This suggests that incorrect responses made during learning are retained in some form and create interference for the correct answer. Importantly, the additional statistics from the online VALMT enabled a comparison of forgetting rates when errors, and therefore the proposed interference, are equated by looking at recall for pairs which encounter no errors (errorless learning). This analysis found no evidence of any difference in forgetting rates for slow and fast learning older participants between learning and recall at 55min, or between recall at 55min and 24hrs. This indicates that the difference in observed recall performance between these groups does not reflect a difference in their underlying forgetting rate, but must instead be caused by some other factor such as interference.

Taken together, the evidence suggests the following explanation for the observed performance of the slow learning older participants: some older participants are suffering from a form of cognitive deficit, perhaps indicative of the early stages of a dementia

process such as Alzheimer's disease, which causes them to make more errors during learning, and those errors create interference for the relevant pairs and result in poorer recall which is apparent at 55min and beyond.

The important relationship between learning errors and subsequent recall identified in this thesis has implications for other researchers. When using any form of learning to criterion procedure, previous published work has only analysed the number of times an entire stimuli set is presented. The total number of individual per-item errors made is not reported or analysed. It may be that if this was recorded and analysed other patterns may become apparent. Most importantly, subtle learning deficits may be identified which will then challenge the original conceptualisation of ALF as a form of late onset accelerated forgetting in those whose learning in unimpaired.

The possibility that VALMT learning performance may act as a marker for risk of developing dementia has important clinical implications given the significant prevalence and impact of the disease and the benefits of early diagnosis. More sensitive cognitive tests will be beneficial to help track behavioral changes during clinical trials (Sperling et al., 2011) and to monitor the impact of treatments once these become available, especially if, as is the case with lecanemab (Alzheimer's Society (2022), these treatments are most effective during the early stage of the disease. Even when biomarker tests such as blood tests are available, it is likely that a combination of biomarker and cognitive tests may prove most effective in early identification. Although a biomarker test may be an important line of evidence when diagnosing a disease such as AD, the wide scale deployment of the equipment and resources needed to administer such tests is many years away, perhaps decades in some regions, so a relatively short cognitive test which can be administered remotely also has significant appeal as a form of triage before administering a more expensive biomarker test. It will therefore be important to validate the ability of VALMT to predict progression to MCI/AD or other forms of dementia, and to clarify whether learning performance alone is a sufficient criteria. If it is, then a targeted form of the test consisting only of the learning stage could be administered within 10-15mins, without the need to return to a website for any delayed testing, which will further simplify administration. Finally, although a biomarker test may provide important evidence when diagnosing an underlying disease such as AD, is it important to remember that what is most important to patients are their cognitive symptoms, which a biomarker test cannot measure.

## 6.3 Does ALF in ageing reflect decay or interference?

There is a long running debate in the memory and forgetting literature about whether forgetting is caused by the passive decay of memory traces, or by memory traces competing or interfering with each other (McGeoch, 1932). As with many such debates, it is likely that both processes are at work, but that one or the other may dominate depending on the duration of the delay between learning and retrieval and other specifics of a given task. The experiments detailed in this thesis can contribute to this ongoing debate.

The results indicate that interference may have contributed to the observed forgetting in two main ways. First, the complex interleaving of learning and testing used in Experiments 1 and 2 may have introduced interference. Second, interference may be created within any single learning period by the learning of other word-pairs and by errors made when learning each pair ('within-learning' interference).

With respect to possible interference introduced by the interleaving of learning and testing, initial analysis of the results from Experiment 2 did not find any strong relationship between the amount of potential interference pairs were exposed to and their subsequent recall at a delay of 55 minutes. However, this evidence was not conclusive, and to help clarify any impact later online experiments removed the complex interleaving and used a single learning period with no other interfering tasks included before delayed recall. Smaller group differences were found with this simplified procedure, with the main impact being improved performance by the older groups. While this may simply reflect a sampling issue, with the online samples being higher performers, it may also indicate that older participants were impacted by interference in Experiments 1 and 2, leading to lower delayed recall and faster forgetting.

With respect to within-learning interference, analysis of the relationship between errors made during learning and subsequent delayed recall showed a common pattern across all experiments and ages. The number of errors made was the strongest predictor of subsequent delayed recall performance. Importantly, it was the specific word-pairs which encountered the most errors which were then recalled most poorly. If a participant makes an error in cued-recall during learning, as used in VALMT, they will have recalled an incorrect second word, which creates a new competitive memory trace; there will now be multiple memory traces, pairing the same first-word with different second words. And

if some participants have a learning deficit which causes them to make more errors then they will create more such competitive traces. Furthermore, the learning to criterion approach may increase the total number of such traces if it forces these slower learning participants to make additional errors whilst attempting to reach criterion, in comparison to their faster learning peers. Discrimination between traces after a delay is more likely to fail as their similarity increases and as the number of similar traces increases (e.g. Poch et al., 2020). The type of errors made during a cued-recall learning-to-criterion procedure can therefore be expected to create interference which may have a significant impact on subsequent recall.

The experiments reported in this thesis have shown a clear pattern for older participants to make more errors during learning, which means they will be exposed to more interference. Any impact from this would be further exacerbated if older people are also more vulnerable to this interference. The design of the current experiments means it was not possible to tease apart whether the poorer delayed recall performance of older groups is simply due to the greater interference they experience, or whether they are also more vulnerable to such interference. However, there is some reason to believe they may be more susceptible. Inhibitory deficit theory (Hasher & Zacks, 1988) proposes that older individuals are less able to control the contents of working memory, leading to increased impact of irrelevant interfering information and resulting in poorer long-term memory performance. Evidence for this comes from studies that introduce distractor stimuli or tasks during the encoding phase (e.g. Mund et al. (2012) and which typically find a greater decrease in subsequent delayed recall in older participants (refer to section 3.1 for more detail). Relating this directly to the VALMT, when a participant makes a learning error they are shown the correct pairing again for two seconds to help them learn the correct association; however, the incorrect answer will still be present in working memory and in older participants this may lead to an impaired representation of the correct pairing, eventually leading to greater forgetting.

Sadeh et al. (2016) propose a 'representation theory of forgetting', in which both decay and interference impact memory, but the primary cause of forgetting depends on the initial type of memory representation. They argue that recollection based traces are dependent mainly on the hippocampus and are impacted more by decay than by interference, while familiarity based traces are created in cortical areas outside the hippocampus and are impacted more by interference than decay. This theory fits with the results detailed in this thesis if the reason older participants, especially the slow learners,

make more errors is that they have a deficit in hippocampal function. This would make it harder for them to create strong recollection based traces, leading to poorer cued-recall learning performance, and cause them to rely more on regions outside the hippocampus which are more prone to interference. If this was the case these slower learners would display both increased learning errors and increased susceptibility to interference. However, without further study such an explanation must remain speculative.

With respect to the possible role of trace decay, the best evidence from the current experiments comes from between-group comparisons of the forgetting curves for errorless pairs, where there will be minimal interference. Forgetting in the absence of any learning errors may be more dependent on a decay process, and thus provides the best indicator of the role of decay. These comparisons found that older participants have a similar underlying forgetting rate to younger participants between completion of learning to criterion and recall at 55min, and may have a slightly faster forgetting rate between 55min and 24hr, which may indicate a faster rate of decay for older participants over longer timeframes. This is consistent with the results of Huppert and Kopelman (1989) who found accelerated forgetting in older participants between 10min and 24hr. However, importantly, there was no evidence that slow learning older participants forgot errorless pairs at a faster rate than faster learning older participants. This suggests that a difference in rate of decay cannot be the explanation for the observed differences in delayed recall performance between these older groups. Instead, these differences must be caused by some other factor such as interference.

Overall, in the current experiments the accelerated forgetting seen in older participants, especially the slow learners, appears to primarily reflect a learning deficit which causes them to make errors and therefore suffer greater interference, rather than reflecting a difference in an underlying decay process. This has implications for any research which has used a learning procedure which introduces errors. The majority of the recent ALF literature has used a learning to criterion procedure in which participants are tested after each learning trial. Any study of this type will introduce increased errors, and therefore greater interference, for anyone who has a learning deficit. If this is the case then the accelerated forgetting observed in many ALF studies may be, at least partly, an artefact of the learning process, rather than reflecting purely a difference in retention. In the case of VALMT, the total number of errors made is recorded and reported, providing a granular measure of learning performance. In contrast, most other work reports only the number of times an entire stimulus set is presented, which is a much coarser measure, and

is likely to fail to spot subtle differences in learning performance; reporting errors in the same way as VALMT would be an improvement. However, if the desire is to measure underlying forgetting in the absence of differential interference then there may be benefit in also testing participants with an errorless learning procedure, perhaps similar to that used by Huppert and Piercy (1977). If ALF can be detected using both approaches that would provide a stronger justification for ALF existing as a genuine phenomenon, rather than being a methodological artefact.

## 6.4 Timeframe for onset of forgetting in ALF

The current research indicates that slow learning older participants display ALF with an early onset, prior to 55mins, but do not display any additional late-onset ALF (beyond 55mins). Previous work investigating ALF in ageing or MCI/AD has typically tested at 30mins, followed by one extended delay far beyond 55mins. For example, Manes at al. (2008) tested at 30min and 6 weeks; Mary et al. (2013) tested at 30min then 7 days; Weston et al. (2018) tested at 30min and 7 days. One exception to this is Baddeley et al. (2014) who tested immediately followed by testing subsets of participants at various intermediate delays of 1 to 24 days, and finally tested all participants 6 weeks. They found evidence of ALF at 6 weeks, but not at the shorter delays. However, the sample size at each intermediate delay was small (8 per delay); it is possible that if larger sample sizes had been used they have found an effect at shorter delays. In summary, the current results are compatible with the work of others, but suggest that if these researchers had tested at a shorter extended delay, within the first few hours after learning, and with more sensitive tests they may have detected ALF within a shorter timeframe.

With reference to the Memory Phases Framework of Radvansky et al. (2022), the ALF detected in the slow learning older participants in the current research appears to start within the Early Long-Term Memory phase (eLTM; 60 seconds to 12 hours) when forgetting should be slowing as hippocampal consolidation occurs. This suggests that it is caused by localised consolidation issues within the medial temporal lobe, rather than issues with transfer of information to longer term storage areas in the neocortex which occurs later (Transitional Long-Term Memory, tLTM; 12hrs to 7 days). However, it remains possible that there may also be older participants who show normal forgetting to 24hr as measured by the VALMT but would then go on to display greater forgetting

during the later stages of the tLTM phase, or during the Long-Lasting Memory phase (LLM; beyond 7 days).

Results also indicate that the main driver of the forgetting displayed by slow learning older participants is interference from the errors made during learning, rather than their underlying forgetting rate. So these participants have a learning deficit, which leads to accelerated forgetting which is apparent by 55min. This raises the possibility that different types of forgetting are being detected by different researchers and in different populations, at least for verbal material. Forgetting which accelerates over a relatively short timeframe, within the first hours after learning (eLTM phase), may indicate forgetting caused by interference in otherwise healthy older participants, or by disrupted hippocampal consolidation in TLE patients (e.g. Jansari & McGibbon 2010; Cassel et al., 2016), and can be detected with current VALMT delay of 55min. In contrast, forgetting which accelerates after days or weeks (tLTM, LLM phases) may indicate issues with transfer of information to, or retention in, longer term storage areas in the neocortex, and its detection would require either an adjusted VALMT to test at these extended delays or the use of another test such as the Crimes test (Baddeley at al., 2019). However, topics raised in this thesis including the importance of interference from learning errors and other methodological issues such as retrieval practice mean paradigms for testing at such extended delays need to be carefully analysed to rule out the possibility that forgetting begins earlier or reflects a subtle learning deficit. It also remains possible that a suitably adjusted VALMT may able to detect all such variants: the learning performance measure can detect subtle learning deficits, the shorter delay can detect early onset forgetting, and a second delay set to 1 week or longer could detect true late onset ALF.

The question of when ALF starts and whether different types of forgetting are being detected by different researchers is also relevant to Weston et al.'s (2018) proposal of a three stage sequence for progressive memory impairment in AD. They suggest that their own study which detected ALF at one week in presymptomatic autosomal dominant (familial) AD indicates that the earliest presymptomatic stage of the disease leads to impaired long-term retention, with encoding and short-term retention remaining normal. The next stage is MCI, in which short-term retention is also impaired but encoding is still normal, and the final stage, clinical AD, impairs encoding too. For evidence that encoding is normal in MCI they cite Ally et al. (2013) who used an image recognition paradigm to show evidence of intact encoding in MCI, tested at approximately 20secs, followed by rapid forgetting visible at approximately 1 minute, while their AD group

showed poor encoding. However, forgetting within 1min and at 1 week reflect very different timeframes, and are likely to involve different neurobiological processes. The one minute interval maps to the Working Memory and Early Long-Term Memory (eLTM) stages of the Radvansky et al. (2022) Memory Phases Framework, which are dependent on active cortical representations and hippocampal consolidation respectively. The one week interval maps to the Transitional Long-Term Memory (tLTM) stage which is reliant on cortical consolidation, the early stage of system, rather than synaptic, consolidation. The proposed 3 stage progression would, therefore, require the impact of AD to change as the disease progresses. The idea of a smooth temporal sequence of progressive symptoms apparent over increasingly shorter delays is appealing, starting with long-term retention, followed by short-term retention and finally encoding. However, Weston et al. provide no explanation for why this specific sequence should occur and given the different processes involved at each timeframe it is not clear why memory processes should be impacted in this order. Nevertheless, if their proposed sequence is correct then the VALMT can potentially detect impairment at all stages; learning performance can provide a measure of encoding (indicating AD), recall at a delay of 30 or 55min can provide a measure of short-term retention (indicating MCI), and recall at a long delay can provide a measure of long-term retention (indicating presymptomatic AD).

The experiments reported in this thesis do not provide any data on forgetting over the one week delay used by Weston et al. (2018). The longest delay used was 24hrs, and no significant ALF between 55min and 24hr was found. Future testing with the second delay changed to one week would be required to check for the presence of forgetting at such extended delays. VALMT experiments did find evidence of ALF at 55min in slow learning older participants. However, these slow learners may have an encoding deficit (since they take more trials to complete learning), so their forgetting would not map to the either the presymptomatic or MCI stages of Weston et al.'s (2018) proposed sequence. In fact, across all the reported VALMT studies there is no evidence for the presence of ALF without impaired learning; ALF was only detected in the slower learners. The strongest evidence for this comes from analysis of forgetting for errorless pairs in the online VALMT experiments; for these pairs there was no evidence of accelerated forgetting over the first 55min or the 55min-24hrs interval.

However, interpretation is complicated by the relatively short timeframe for forgetting in MCI found by Alay et al. (2013) and cited by Weston et al. (2018). They found

evidence of forgetting at one minute in their MCI group. Such a short delay is less than the total time for one presentation or test of the entire stimulus set in many verbal tests (short story, word list, VALMT word-pair set). This means such rapid forgetting would impact the apparent learning between trials of an entire stimulus set and show up as a learning deficit, and such tests will therefore be unable to distinguish an encoding deficit from intact encoding followed by forgetting within 1 minute. It is therefore possible than some of the slower learners in VALMT studies actually had intact encoding followed by very rapid forgetting, which would be indicative of MCI according to the Weston et al. model. If this is the case, then the VALMT data would fit their model, and further suggest that although no participants in the current research reported any diagnosis of MCI some may have reached that stage but remain undiagnosed.

A final possible explanation for the lack of ALF observed between 55min and 24hrs in the older groups in VALMT experiments is that the samples did not include any participants with preclinical AD. While this is possible, the sample sizes (e.g. Expt 4, 32 aged 60yr+) combined with the prevalence of AD and MCI in the UK population (Richardson et al., 2019: aMCI prevalence 15.2% for 65yrs and over) makes this unlikely.

Overall, while the VALMT data may be compatible with the Weston et al. (2018) model for the progression of AD symptoms, this is not conclusive and further testing will be required, particularly at the extended 1 week delay.

## 6.5 Memory performance and prevalence of ALF in the general population

Analysis of memory across the lifespan showed a trend for performance to start deteriorating (recall scores decreasing; learning errors, memory complaints and forgetting rate increasing) in the 57-69 age band. Comparing this to previous lifespan studies investigating forgetting over extended delays of at least 24hrs, Davis et al. (2003) found evidence of accelerated forgetting after 24hr in their oldest age band (76-90yrs), while Huppert and Kopelman (1989) found accelerated forgetting in their 38-64yr middle-aged group. The data from Experiment 4 therefore indicates an onset of decline at an age more consistent with Huppert and Kopelman. However, larger samples will be required to more precisely identify when this deterioration starts. At the other end of the age range, there was no evidence of any lower performance in the 16-17yr age range, so it seems that

memory performance as measured by the VALMT has already reached its peak level by this age, which is also consistent with Huppert and Kopelman.

To investigate the prevalence of ALF within the general population it will be necessary to define a criterion for diagnosis. The experiments covered in this thesis have shown that forgetting rates in the general population have continuous distributions, with no clear-cut boundary between those that could be classed as normal or displaying ALF. In this case it becomes necessary to select a statistical criterion. For example, we may choose to diagnose those 2.5% or 5% of the population with the fastest forgetting rates as displaying ALF. It will then be possible to analyse the frequency of ALF based on demographic variables such as age or gender. However, since such analyses will be working with only 2.5% or 5% of the overall sample, such analyses will need to wait until data from a much larger number of participants has been gathered.

## 6.6 Conclusion

The current programme of research evaluated use of the Verbal Associative Learning and Memory Test (VALMT) paradigm to investigate learning and forgetting over both short and long delays, with a focus on accelerated forgetting rates.

VALMT proved to be more sensitive than an existing clinical test (WMS-LM). It can be administered both face-to-face, or using a new online version developed as part of this research. The online version works successfully with all age groups, and makes VALMT easy to administer at scale without the need for travel and training of test administrators.

Early experiments showed presence of ceiling effects. A detailed analysis of the impact of such effects in the type of experimental design used in forgetting research led to the development of a set of practical guidelines for use by researchers within the field. It was possible to adjust the VALMT design such that results complied with these guidelines, reducing ceiling effects to a level that will not influence results while still allowing memory to be tested at both short and long delays in groups with a variety of performance levels.

VALMT provides granular data which can be used when building and evaluating models and theories of memory and forgetting and, importantly, how these processes are impacted by disease processes such as Alzheimer's disease.

Perhaps most importantly, VALMT identified accelerated forgetting within 55min for older participants who learn more slowly. There is potential for this learning deficit and

subsequent forgetting to be used to identify older individuals who are at risk of developing dementia, which will be of significant benefit in research and clinical practice.

# Appendix A Experiment 1 results summary

## A.1 Participants

Two groups of participants were assessed: 43 Younger participants aged 20-30 (24F, 19M: Mean Age: 22.84, SD: 2.46) were compared to 26 Older participants aged 65-80 (16F, 10M: Mean Age: 70.62, SD: 4.71). Due to a technical error learning performance data was unavailable for 7 Older participants. Analysis with sub-groups based on learning performance analysis was therefore conducted with 19 Older participants.

## A.2 Procedure and materials

The VALMT procedure and materials were the same as those detailed in the main study, except that materials and instructions were translated into Romanian. Unlike the main study, due to limits on time and resources no standard clinical tests were performed, and no data was gathered on subjective memory complaints or sleep patterns.

## A.3 Results

Figure A-1 shows the delayed recall performance of the Younger group and the combined Older group. Figure A-2 shows the performance with Older group split into fast and slow learners based on learning trials required to reach criterion (Fast Older, Slow Older), implemented due to an apparent bimodal distribution in this variable.

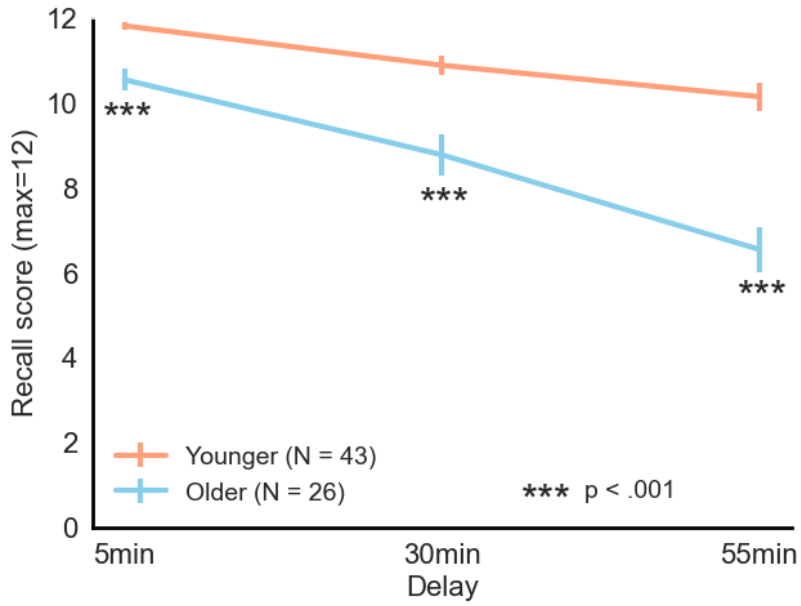


*Figure A-1 Pilot study: Mean VALMT recall scores as a function of time delay and group (error bars +/- 1SE)*

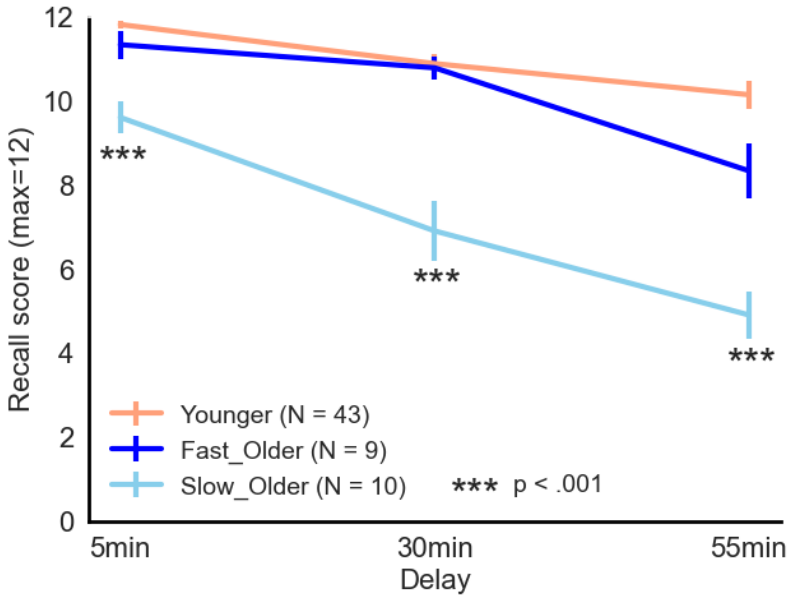


*Figure A-2 Pilot study: Mean VALMT recall scores as a function time delay and group, separating the Older group into two groups based on initial learning (error bars represent one standard error).*

### A.3.1 Combined Older group compared to Younger group

A mixed factors ANOVA with within-subjects factor Delay (5min vs 30min vs 55min) and between-subjects factor Group (Younger vs Older) was used to analyse cued recall performance across all delay intervals. Significant main effects of Delay ($F(2, 121) = 69.08$, $p < .001$, $\eta_p^2 = 0.51$, $BF_{10} = 3.34*10^{14}$) and Group ($F(1, 67) = 40.71$, $p < .001$, $\eta_p^2 = 0.34$, $BF_{10} = 2.84*10^{8}$) were found, along with a significant interaction ($F(2, 121) = 11.89$, $p < .001$, $\eta_p^2 = 0.15$, $BF_{10} = 3100$).

Independent samples t-tests were used to compare recall performance between groups at each delay interval. The Older group scored significantly lower than the Younger group at all 3 delay intervals; 5mins ($M_{Older} = 10.54$ pairs, $M_{Younger} = 11.81$ pairs; $t(31.2) = 4.71$, $p < .001$, $d = 1.38$, $BF_{10} = 25511$), 30mins ($M_{Older} = 8.77$ pairs, $M_{Younger} = 10.88$ pairs; $t(35.8) = 3.91$, $p < .001$, $d = 1.09$, $BF_{10} = 519$) and 55mins ($M_{Older} = 6.54$ pairs, $M_{Younger} = 10.14$ pairs; $t(67) = 6.07$, $p < .001$, $d = 1.51$, $BF_{10} = 1.20*10^{5}$).

Forgetting rates were calculated as amount of information lost between two consecutive time points relative to the amount that had been recalled at the earlier of the two time points. Therefore, the 'early' forgetting rate (that between the 5 and 30 minute time points) was calculated as [5min score – 30min score]/5min score, and the 'late' forgetting rate (that between the 30 and 55 minute time points) was calculated as [30min

score – 55min score] / 30min score. Independent samples t-tests found the Older group had a significantly greater early-forgetting rate ($M_{Older}$ = .17, $M_{Younger}$ = .08; $t(34.3)$ = 2.15, $p$ = .039, $d$ = 0.61, $BF_{10}$ = 3.11) and late forgetting rate ($M_{Older}$ = .25, $M_{Younger}$ = .06; $t(67)$ = 3.53, $p$ = .001, $d$ = 0.88, $BF_{10}$ = 38.38).

The analysis above uses all Older participants ($N$ = 26), including those for whom learning data was not available. For comparison, the analysis was repeated using only those Older participants for whom learning data was available ($N$ = 19). All tests of statistical significance produced the same result using this smaller group.

### A.3.2 Fast and slow learning Older groups

A mixed factors ANOVA with within-subjects factor Delay (5min vs 30min vs 55min) and between-subjects factor Group (Younger vs Fast_Older vs Slow_Older) identified significant main effects of Delay ($F(2,118)$ = 56.94, $p$ < .001, $\eta_p^2$ = 0.49, $BF_{10}$ = 2.31*10$^{14}$) and Group ($F(2,59)$ = 48.08, $p$ < .001, $\eta_p^2$ = 0.60, $BF_{10}$ = 7.17*10$^{9}$), and a significant interaction ($F(4,118)$ = 6.93, $p$ < .001, $\eta_p^2$ = 0.19, $BF_{10}$ = 1337). Bonferroni post hoc tests found no significant difference between the Younger and Fast_Older group ($p$ = .195, $BF_{10}$ = 1.83), but significant differences between the Younger and Slow_Older ($p$ < .001, $BF_{10}$ = 1.61*10$^{15}$) and importantly, between the Fast_Older and Slow_Older ($p$ < .001, $BF_{10}$ = 2383).

Recall scores at each delay were compared using one-way ANOVAs with Bonferroni post hoc tests of significant results. There was a significant difference between the means at all 3 delays (5mins: $F(2,59)$ = 33.63, $p$ < .001, $\eta_p^2$ = 0.53, $BF_{10}$ = 4.45*10$^{7}$; 30mins: $F(2,59)$ = 27.25, $p$ < .001, $\eta_p^2$ = 0.48, $BF_{10}$ = 1.96*10$^{6}$; 55mins: $F(2,59)$ = 26.46, $p$ < .001, $\eta_p^2$ = 0.47, $BF_{10}$ = 1.85*10$^{6}$). At all delays the Slow_Older group performed statistically below the Younger (5min: $p$ < .001, $BF_{10}$ = 3.76*10$^{8}$; 30min, $p$ < .001, $BF_{10}$ = 9.80*10$^{5}$; 55min: $p$ < .001, $BF_{10}$ = 1.95*10$^{6}$) and Fast_Older groups (5min: $p$ < .001, $BF_{10}$ = 12.39; 30min, $p$ < .001, $BF_{10}$ = 151; 55min: $p$ = .002, $BF_{10}$ = 32.48). There was no significant difference between the Younger and Fast_older groups' performance at any delay (5min: $p$ = .281, $BF_{10}$ = 1.43$^{6}$; 30min, $p$ = 1.00, $BF_{10}$ = 0.35; 55min: $p$ = 0.063, $BF_{10}$ = 2.56).

Forgetting rates were compared across groups using one-way ANOVAs. The difference between the means approached significance for early forgetting (Welch's $F(2,13.6)$ = 3.68 $p$ = .053, $\eta_p^2$ = 0.24, $BF_{10}$ = 65.02) and was significant for late forgetting

($F(2,59) = 5.71$, $p = 0.005$, $\eta_p^2 = 0.16$, $BF_{10} = 10.83$). Bonferroni post hoc tests found the Slow_Older group had faster early and late forgetting rates than the Younger group (early: $p = .001$, $BF_{10} = 93.74$; late: $p = .013$, $BF_{10} = 7.88$), and higher early, but not late, forgetting rates than the Fast_Older group (early: $p = .002$, $BF_{10} = 3.60$; late: $p = 1.00$, $BF_{10} = 0.46$). There was no significant difference between the Younger and Fast_Older groups' early or late forgetting rates (early: $p = 1.00$, $BF_{10} = 0.47$; late: $p = .108$, $BF_{10} = 1.81$).

# Appendix B Word-pair stimuli sets

All words, including recognition test foils, were selected from an online lexical database which provides data on parameters including frequency, imageability and concreteness: https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

## B.1 Word-pairs used in Experiment 2

| List 1 | List 2 | List 3 |
|---|---|---|
| L1a<br>flag, grass<br>whale, thief<br>paste, yacht<br>spoon, nerve | L2a<br>crown, paint<br>fruit, gang<br>track, sheet<br>rail, boss | L3a<br>bench, skirt<br>purse, cook<br>swim, green<br>earth, queen |
| L1b<br>sweat, slope<br>globe, suite<br>shark, cross<br>vest, pork | L2b<br>duck, fork<br>dinner, doctor<br>sole, leak<br>bark, sail | L3b<br>lock, beef<br>waist, snake<br>brain, scale<br>corn, nail |
| L1c<br>fight, clock<br>coil, moth<br>judge, birth<br>skull, juice | L2c<br>dance, sheep<br>blood, sound<br>guest, plane<br>weed, clay | L3c<br>tribe, watch<br>well, dime<br>shirt, chief<br>vein, tune |

## B.2 Word-pairs used in Online VALMT Experiments 3, 4, 5, 6 and 10 (12 pair set)

| List 1 |
|---|
| waist, spray |
| stove, dirt |
| grip, coil |
| juice, pork |
| sole, crow |
| blade, calf |
| fork, vest |
| steel, bark |
| pump, loop |
| troop, leak |
| tribe, sweep |
| rash, paste |

## B.3 Word-pairs used in Online VALMT Experiment 7 (16 pair set)

Conditions:

1. Learnt to 2-recall criterion, tested at 30mins

2. Learnt to 1-recall criterion, tested at 30mins
3. Learnt to 5-recall criterion, tested at 24hrs
4. Learnt to 3-recall criterion, tested at 24hrs

| List 1 | Condition |
|---|---|
| waist, spray | 1 |
| stove, dirt | 2 |
| grip, coil | 3 |
| juice, pork | 4 |
| sole, crow | 1 |
| blade, calf | 2 |
| fork, vest | 3 |
| steel, bark | 4 |
| pump, loop | 1 |
| troop, leak | 2 |
| tribe, sweep | 3 |
| rash, paste | 4 |
| tune, veal | 1 |
| birth, suite | 2 |
| globe, lock | 3 |
| thief, vein | 4 |

## B.4 Word-pairs used in Online VALMT Experiments 8 and 9 (16 pair set)

Conditions:

1. Learnt to 1-recall criterion, tested at 30mins
2. Learnt to 3-recall criterion, tested at 24hrs

| List 1 | Condition |
|---|---|
| waist, spray | 1 |
| stove, dirt | 1 |
| grip, coil | 2 |
| juice, pork | 2 |
| sole, crow | 1 |
| blade, calf | 1 |
| fork, vest | 2 |
| steel, bark | 2 |
| pump, loop | 1 |
| troop, leak | 1 |
| tribe, sweep | 2 |
| rash, paste | 2 |
| tune, veal | 1 |
| birth, suite | 1 |
| globe, lock | 2 |
| thief, vein | 2 |

## B.5 Words used in Online VALMT recognition testing in Experiment 10 (12 pair set)

Conditions highlighted in colour:

1. Target, correct answer
2. Second word from another pair, so previously seen
3. Unseen foil

| Cue Word | FCR Choice 1 | FCR Choice 2 | FCR Choice 3 | FCR Choice 4 |
|---|---|---|---|---|
| waist | dirt | crumb | spray | crook |
| stove | bark | flare | cast | dirt |
| grip | speck | coil | calf | heap |
| juice | pork | scale | lane | sweep |
| sole | spray | guide | crow | beam |
| blade | pint | chart | pork | calf |
| fork | flash | vest | leak | nerve |
| steel | bark | lace | pile | paste |
| pump | loop | crow | link | lump |
| troop | coil | wink | prize | leak |
| tribe | loop | brake | sweep | track |
| rash | paste | bump | port | vest |

# Appendix C VALMT screen captures

Figure 9.1 shows the screen displayed during the initial presentation of each pair during learning. Figure 9.2 shows the screen displayed during cued-recall testing, during both learning and delayed recall. Figure 9.3 and 9.4 show the screen displayed after a correct or incorrect cued-recall response respectively during the learning phase. Figure 9.5 shows the screen displayed during recognition testing. No feedback is provided after responses during delayed recall or recognition testing.

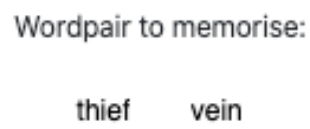


*Figure C-1 Memorisation screen*

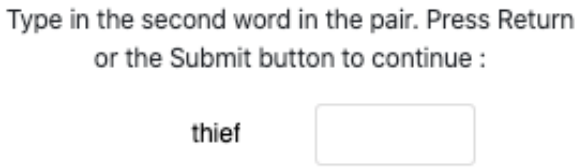


*Figure C-2 Cued-recall test screen*

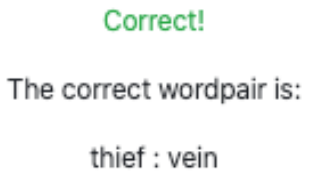


*Figure C-3 Feedback screen after correct cued-recall response*

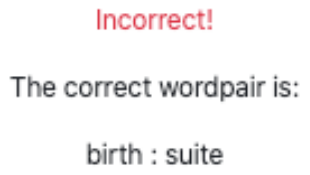


*Figure C-4 Feedback screen after incorrect cued-recall response*

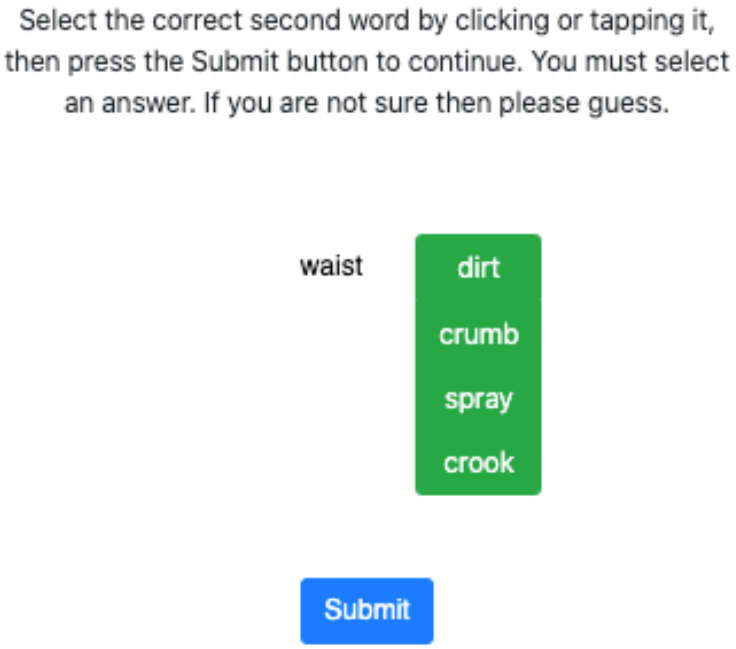


*Figure C-5 Recognition test screen*

# References


Albert, M. S. (2008). The neuropsychology of the development of Alzheimer's disease. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (pp. 97–132). Psychology Press. https://www.routledgehandbooks.com/doi/10.4324/9780203837665.ch3

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B. & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 270-279. https://doi.org/10.1016/j.jalz.2011.03.008

Ally, B. A., Hussey, E. P., Ko, P. C., & Molitor, R. J. (2013). Pattern separation and pattern completion in Alzheimer's disease: evidence of rapid forgetting in amnestic mild cognitive impairment. *Hippocampus*, *23*(12), 1246-1258. https://doi.org/10.1002/hipo.22162

Alvarez, P. & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences USA, 91*, 7041-7045. https://doi.org/10.1073/pnas.91.15.7041

Alzheimer's Association. (2018). 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia, 14(3)*, 367-429.

Alzheimer's Society (2022, Dec 14). *Lecanemab: A new drug for early-stage Alzheimer's disease***.** https://www.alzheimers.org.uk/blog/lecanemab-new-drug-early-stage-alzheimers-disease

Atherton, K. E., Nobre, A. C., Zeman, A. Z., & Butler, C. R. (2014). Sleep-dependent memory consolidation and accelerated forgetting. *Cortex, 54*, 92–105. https://doi.org/10.1016/j.cortex.2014.02.009

Baddeley, A. D., Atkinson, A. L., Hitch, G. J., & Allen, R. J. (2021). Detecting accelerated long-term forgetting: A problem and some solutions. *Cortex*, *142*, 237-251. https://doi.org/10.1016/j.cortex.2021.03.038

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. https://doi.org/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 ageing and the Crimes Test. *Memory, 22*(8), 1052–1059. https://doi.org/10.1080/09658211.2013.865753

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*(1), 53-68. https://doi.org/10.1016/0028-3932(94)90068-X

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in psychology*, *7*, 1378. https://doi.org/10.3389/fpsyg.2016.01378

Braak, H., Braak, E. (1998). Evolution of neuronal changes in the course of Alzheimer's disease. In: Jellinger, K., Fazekas, F., Windisch, M. (eds) Ageing and Dementia. Journal of Neural Transmission. Supplementa, vol 53. Springer, Vienna. https://doi.org/10.1007/978-3-7091-6467-9_11

Breen, R. (1996). Regression models: Censored, sample selected, or truncated data. Thousand Oaks, CA: Sage.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 133–139). American Psychological Association. https://doi.org/10.1037/14805-009

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286-291. https://doi.org/10.3758/BRM.42.1.286

Butler, C. R., Graham, K. S., Hodges, J. R., Kapur, N., Wardlaw, J. M., & Zeman, A. Z. J. (2007). The syndrome of transient epileptic amnesia. *Annals of Neurology, 61*(6), 587–598. https://doi.org/10.1002/ana.21111

Buysse, D. J., Reynolds III, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry research*, *28*(2), 193-213. https://doi.org/10.1016/0165-1781(89)90047-4

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

Cassel, A., & Kopelman, M. D. (2019). Have we forgotten about forgetting? A critical review of 'accelerated long-term forgetting' in temporal lobe epilepsy. *Cortex*, *110*, 141-149. https://doi.org/10.1016/j.cortex.2017.12.012

Christensen, H., Kopelman, M. D., Stanhope, N., Lorentz, L., & Owen, P. (1998). Rates of forgetting in Alzheimer dementia. *Neuropsychologia*, *36*(6), 547-557. https://doi.org/10.1016/S0028-3932(97)00116-4

Clare, L., Wilson, B. A., Carter, G., Roth, I., & Hodges, J. R. (2002). Relearning face-name associations in early Alzheimer's disease. *Neuropsychology*, *16*(4), 538. https://psycnet.apa.org/doi/10.1037/0894-4105.16.4.538

Coupé, P., Manjón, J. V., Lanuza, E., & Catheline, G. (2019). Lifespan changes of the human brain in Alzheimer's disease. *Scientific reports*, *9*(1), 1-12. https://doi.org/10.1038/s41598-019-39809-8

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. Trends in Cognitive Sciences, 10, 131–138. https://doi.org/10.1016/j.tics.2006.01.007

Davis, H. P., Small, S. A., Stern, Y., Mayeux, R., Feldstein, S. N., & Keller, F. R. (2003). Acquisition, recall, and forgetting of verbal information in long-term memory by young, middle-aged, and elderly individuals. *Cortex*, *39*(4-5), 1063-1091. https://doi.org/10.1016/S0010-9452(08)70878-5

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). California verbal learning test research edition manual. *San Antonio: The Psychological Corporation*.

de Rover, M., Pironti, V. A., McCabe, J. A., Acosta-Cabronero, J., Arana, F. S., Morein-Zamir, S., ... & Sahakian, B. J. (2011). Hippocampal dysfunction in patients with mild cognitive impairment: a functional neuroimaging study of a visuospatial paired associates learning task. *Neuropsychologia*, *49*(7), 2060-2070. https://doi.org/10.1016/j.neuropsychologia.2011.03.037

Diana, R.A., Reder, L.M., Arndt, J. *et al.* Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review* **13**, 1–21 (2006). https://doi.org/10.3758/BF03193807

Dudai, Y. (2004). The Neurobiology of Consolidations, Or, How Stable is the Engram? *Annual Review of Psychology, 55*(1), 51–86. https://doi.org/10.1146/annurev.psych.55.090902.142050

Dunne, R. A., Aarsland, D., O'Brien, J. T., Ballard, C., Banerjee, S., Fox, N. C., ... & Burns, A. (2021). Mild cognitive impairment: the Manchester consensus. *Age and ageing*, *50*(1), 72-80. https://doi.org/10.1093/ageing/afaa228

Ebbinghaus, H. (1885) *Memory: A Contribution to Experimental Psychology* Refurbished 1964 New York: Dover

Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., & Thompson-Schill, S. L. (2006). Interfering with Theories of Sleep and Memory: Sleep, Declarative Memory, and Associative Interference. *Current Biology, 16*(13), 1290–1294. https://doi.org/10.1016/j.cub.2006.05.024

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

Fandakova, Y., Lindenberger, U., & Shing, Y. L. (2015). Episodic memory across the lifespan: general trajectories and modifers. The Wiley handbook on the cognitive neuroscience of memory (pp. 309–325). Wiley.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*(6958), 614–616. https://doi.org/10.1038/nature01951

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Fox, N. C., Warrington, E. K., Seiffer, A. L., Agnew, S. K., & Rossor, M. N. (1998). Presymptomatic cognitive deficits in individuals at risk of familial Alzheimer's disease. A longitudinal prospective study. *Brain: a journal of neurology*, *121*(9), 1631-1639. https://doi.org/10.1093/brain/121.9.1631

Ganor-Stern, D., Seamon, J. G., & Carrasco, M. (1998). The role of attention and study time in explicit and implicit memory for unfamiliar visual stimuli. *Memory & Cognition, 26*(6), 1187–1195. https://doi.org/10.3758/BF03201194

Gardiner, J. M. & Java, R. I. (1993). Recognising and remembering. In A. Collins, S. Gathercole, M. Conway & P. Morris (Eds), *Theories of memory* (pp. 163-188). Hillsdale, NJ: Erlbaum.

Gascoigne, M. B., Barton, B., Webster, R., Gill, D., Antony, J., & Lah, S. S. (2012). Accelerated long-term forgetting in children with idiopathic generalized epilepsy: Accelerated Long-Term Forgetting in Children. *Epilepsia, 53*(12), 2135–2140. https://doi.org/10.1111/j.1528-1167.2012.03719.x

Giambra, L. M., & Arenberg, D. (1993). Adult age differences in forgetting sentences. Psychology and Aging, 8(3), 451–462. https://doi.org/10.1037//0882-7974.8.3.451

Green, R. E., & Kopelman, M. D. (2002). Contribution of recollection and familiarity judgements to rate of forgetting in organic amnesia. *Cortex*, *38*(2), 161-178. https://doi.org/10.1016/S0010-9452(08)70648-8

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of learning and motivation*, *22*, 193-225. https://doi.org/10.1016/S0079-7421(08)60041-9

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9-10), 1306-1336. https://doi.org/10.1080/13506285.2013.823140

Hirnstein, M., Stuebs, J., Moè, A., & Hausmann, M. (2022). Sex/Gender Differences in Verbal Fluency and Verbal-Episodic Memory: A Meta-Analysis. *Perspectives on Psychological Science*, *0*(0). https://doi.org/10.1177/17456916221082116

Hoefeijzers, S., Dewar, M., Della Sala, S., Butler, C., & Zeman, A. (2015). Accelerated long-term forgetting can become apparent within 3–8 hours of wakefulness in patients with transient epileptic amnesia. *Neuropsychology, 29*(1), 117–125. http://dx.doi.org/10.1037/neu0000114

Hoefeijzers, S., Dewar, M., Della Sala, S., Zeman, A., & Butler, C. (2013). Accelerated long-term forgetting in transient epileptic amnesia: An acquisition or consolidation deficit? *Neuropsychologia, 51*(8), 1549–1555. https://doi.org/10.1016/j.neuropsychologia.2013.04.017

Holdnack, H. A. (2001). Wechsler test of adult reading: WTAR. *San Antonio, TX: The Psychological Corporation*.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Hulicka, I. M., & Weiss, R. L. (1965). Age differences in retention as a function of learning. *Journal of Consulting Psychology, 29*(2), 125–129. https://doi.org/10.1037/h0021793

Huppert, F. A., & Kopelman, M. D. (1989). Rates of forgetting in normal ageing: A comparison with dementia. *Neuropsychologia, 27*(6), 849–860. https://doi.org/10.1016/0028-3932(89)90008-0

Huppert, F. A., & Piercy, M. (1977). Recognition memory in amnesic patients: a defect of acquisition?. *Neuropsychologia*, *15*(4-5), 643-652. https://doi.org/10.1016/0028-3932(77)90069-0

Huppert, F. A., & Piercy, M. (1978). Dissociation between learning and remembering in organic amnesia. *Nature*, *275*(5678), 317-318. https://doi.org/10.1038/275317a0

Hussey, E. P., Smolinsky, J. G., Piryatinsky, I., Budson, A. E., & Ally, B. A. (2012). Using Mental Imagery to Improve Memory in Patients With Alzheimer Disease: Trouble

Generating or Remembering the Mind's Eye? *Alzheimer Disease & Associated Disorders, 26*(2), 124–134. https://dx.doi.org/10.1097%2FWAD.0b013e31822e0f73

Isaac, C. L., & Mayes, A. R. (1999)a. Rate of forgetting in amnesia: I. Recall and recognition of prose. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 942–962. https://doi.org/10.1037/0278-7393.25.4.942

Isaac, C. L., & Mayes, A. R. (1999)b. Rate of forgetting in amnesia: II. Recall and recognition of word lists at different levels of organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 963–977. https://doi.org/10.1037/0278-7393.25.4.963

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*(6), 1707–1715. https://doi.org/10.1016/j.neuropsychologia.2010.02.018

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. https://doi.org/10.1006/brcg.1997.0927

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kopelman, M. D., & Stanhope, N. (1997). Rates of forgetting in organic amnesia following temporal lobe, diencephalic, or frontal lobe lesions. *Neuropsychology, 11*(3), 343–356. https://doi.org/10.1037/0894-4105.11.3.343

Kopelman, M. D. (1985). Rates of forgetting in Alzheimer-type dementia and Korsakoff's syndrome. *Neuropsychologia*, *23*(5), 623-638. https://doi.org/10.1016/0028-3932(85)90064-8

Lah, S., Black, C., Gascoigne, M. B., Gott, C., Epps, A., & Parry, L. (2017). Accelerated long-term forgetting is not epilepsy specific: evidence from childhood traumatic brain injury. *Journal of neurotrauma*, *34*(17), 2536-2544. https://doi.org/10.1089/neu.2016.4872

Laverick, T., Evans, S., Freeston, M., & Baddeley, A. (2021). The use of novel measures to detect Accelerated Long-term forgetting in people with epilepsy: The Crimes Test and Four Doors Test. *Cortex*, *141*, 144-155. https://doi.org/10.1016/j.cortex.2021.03.024

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lewis, P., & Kopelman, M. D. (1998). Forgetting rates in neuropsychiatric disorders. *Journal of Neurology, Neurosurgery & Psychiatry*, *65*(6), 890-898.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(2), 397–406. https://doi.org/10.1037/0278-7393.11.2.397

Malpass, R. S., Tredoux, C. G., Compo, N. S., McQuiston-Surrett, D., MacLin, O. H., Zimmerman, L. A., & Topp, L. D. (2008). Study space analysis for policy development. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(6), 789-801. https://doi.org/10.1002/acp.1483

Mameniskiene, R., Jatuzis, D., Kaubrys, G., & Budrys, V. (2006). The decay of memory between delayed and long-term recall in patients with temporal lobe epilepsy. *Epilepsy & Behavior, 8*(1), 278–288. https://doi.org/10.1016/j.yebeh.2005.11.003

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*(7), 1067–1070. https://doi.org/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. https://doi.org/10.3389/fpsyg.2013.00750

Mayes, A. R., Hunkin, N. M., Isaac, C., & Muhlert, N. (2019). Are there distinct forms of accelerated forgetting and, if so, why? *Cortex*, *110*, 115-126. https://doi.org/10.1016/j.cortex.2018.04.005

McBee, M. (2010). Modeling Outcomes With Floor or Ceiling Effects: An Introduction to the Tobit Model. Gifted Child Quarterly, 54(4), 314–320. https://doi.org/10.1177/0016986210379095

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review, 39*(4), 352–370. https://doi.org/10.1037/h0069819

McGibbon, T., & Jansari, A. S. (2013). Detecting the onset of accelerated long-term forgetting: evidence from temporal lobe epilepsy. *Neuropsychologia, 51*(1), 114–122. https://doi.org/10.1016/j.neuropsychologia.2012.11.004

McGibbon, T., Jansari, A., Demirjian, J., Nemes, A., & Opre, A. (2022). Accelerated forgetting in healthy older samples: Implications for methodology, future ageing studies, and early identification of risk of dementia. Quarterly Journal of Experimental Psychology, 0(0). https://doi.org/10.1177/17470218221113412

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., ... & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 263-269. https://doi.org/10.1016/j.jalz.2011.03.005

Morse, C. K. (1993). *Does variability increase with age? An archival study of cognitive measures*. American Psychological Association. Retrieved from http://psycnet.apa.org/journals/pag/8/2/156/

Moscovitch, M. (2008). The hippocampus as a "stupid," domain-specific module: Implications for theories of recent and remote memory, and of imagination. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 62*(1), 62–79. https://doi.org/10.1037/1196-1961.62.1.62

Mitchell, A. J., Beaumont, H., Ferguson, D., Yadegarfar, M., & Stubbs, B. (2014). Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: meta-analysis. *Acta Psychiatrica Scandinavica*, *130*(6), 439-451. https://doi.org/10.1111/acps.12336

Mulhert, N., Milton, F., Butler, C. R., Kapur, N., & Zeman, A. Z. (2010). Accelerated forgetting of real-life events in transient epileptic amnesia. *Neuropsychologia, 48*, 3235-3244. https://doi.org/10.1016/j.neuropsychologia.2010.07.001

Mund, I., Bell, R., & Buchner, A. (2012). Aging and interference in story recall. *Experimental aging research*, *38*(1), 20-41. https://doi.org/10.1080/0361073X.2012.636724

Nadel, L. & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7*, 217-227. https://doi.org/10.1016/S0959-4388(97)80010-4

National Institute for Health and Care Excellence. (2018). Dementia: assessment, management and support for people living with dementia and their carers. *Retrieved from NICE website: https://www.nice.org.uk/guidance/ng97/chapter/Recommendations#diagnosis*

Naveh-Benjamin, M., Hussain, Z., Guez, J., & Bar-On, M.(2003). Adult age differences in episodic memory: further support for an associative-deficit hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(5), 826. https://psycnet.apa.org/doi/10.1037/0278-7393.29.5.826

Noyes, E., Davis, J. P., Petrov, P., Gray, K. L. H., Ritchie, K. (2021). The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science, 8*, 201169. https://doi.org/10.1098/rsos.201169

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411-419. Available at SSRN: https://ssrn.com/abstract=1626226

Park, D. C., Puglisi, J. T., & Smith, A. D. (1986). Memory for pictures: Does an age-related decline exist? *Psychology and Aging, 1*(1), 11–17. https://doi.org/10.1037/0882-7974.1.1.11

Park, D. C., Royal, D., Dudley, W., & Morrell, R. (1988). Forgetting of pictures over a long retention interval in young and older adults. *Psychology and Aging, 3*(1), 94–95. https://doi.org/10.1037/0882-7974.3.1.94

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., Phillips, C., Degueldre, C., Del Fiore, G., Aerts, J., Luxen, A. & Maquet, P. (2004). Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron, 44*, 535-545. https://doi.org/10.1016/j.neuron.2004.10.007

Poch, C., Toledano, R., García-Morales, I., Prieto, A., García-Barragán, N., Aledo-Serrano, Á., Gil-Nagel, A., & Campo, P. (2020). Mnemonic discrimination in patients with unilateral mesial temporal lobe epilepsy relates to similarity and number of events stored in memory. *Neurobiology of Learning and Memory*, *169*, 107177. https://doi.org/10.1016/j.nlm.2020.107177

Radvansky, G. A., Doolen, A. C., Pettijohn, K. A., & Ritchey, M. (2022). A new look at memory retention and forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001110

Ricci, M., Mohamed, A., Savage, G., & Miller, L. A. (2015). Disruption of learning and long-term retention of prose passages in patients with focal epilepsy. *Epilepsy & Behavior*, *51*, 104-111. https://doi.org/10.1016/j.yebeh.2015.06.016

Richardson, C., Stephan, B.C.M., Robinson, L. *et al.* Two-decade change in prevalence of cognitive impairment in the UK. *Eur J Epidemiol* **34**, 1085–1092 (2019). https://doi.org/10.1007/s10654-019-00554-x

Rivera-Lares, K., Logie, R., Baddeley, A., & Della Sala, S. (2022). Rate of forgetting is independent of initial degree of learning. *Memory & cognition*, 1-13. https://doi.org/10.3758/s13421-021-01271-1

Rivera-Lares, K., Sala, S. D., Baddeley, A., & Logie, R. (2023). Rate of forgetting is independent from initial degree of learning across different age groups. *Quarterly Journal of Experimental Psychology*, *76*(7), 1672-1682. https://doi.org/10.1177/17470218221128780

Roediger, H. L., & Smith, M. A. (2012). The "pure-study" learning curve: The learning curve without cumulative testing. *Memory & Cognition, 40*(7), 989–1002. https://doi.org/10.3758/s13421-012-0213-5

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111%2Fj.1467-9280.2006.01693.x

Rybarczyk, B. D., Hart, R. P., & Harkins, S. W. (1987). Age and forgetting rate with pictorial stimuli. *Psychology and Aging, 2*(4), 404–406. https://doi.org/10.1037/0882-7974.2.4.404

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2016). Forgetting Patterns Differentiate Between Two Forms of Memory Representation. *Psychological Science*, *27*(6), 810–820. https://doi.org/10.1177/0956797616638307

Sapkota, R. P., van der Linde, I., Lamichhane, N., Upadhyaya, T., & Pardhan, S. (2017). Patients with mild cognitive impairment show lower visual short-term memory performance in feature binding tasks. *Dementia and geriatric cognitive disorders extra*, *7*(1), 74-86. https://doi.org/10.1159/000455831

Sheldon, S., Amaral, R., & Levine, B. (2017). Individual differences in visual imagery determine how event information is remembered. *Memory, 25*(3), 360–369. https://psycnet.apa.org/doi/10.1080/09658211.2016.1178777

Shing, Y. L., Werkle-Bergner, M., Brehmer, Y., Müller, V., Li, S. C., & Lindenberger, U. (2010). Episodic memory across the lifespan: The contributions of associative and strategic components. *Neuroscience & Biobehavioral Reviews*, *34*(7), 1080-1091. https://doi.org/10.1016/j.neubiorev.2009.11.002

Silva, D., Guerreiro, M., Maroco, J., Santana, I., Rodrigues, A., Marques, J. B., & de Mendonça, A. (2012). Comparison of four verbal memory tests for the diagnosis and predictive value of mild cognitive impairment. *Dementia and geriatric cognitive disorders extra*, *2*(1), 120-131. https://dx.doi.org/10.1159%2F000336224

Slamecka, N. J., & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(3), 384–397. https://doi.org/10.1037/0278-7393.9.3.384

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., & Park, D. C. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 280-292. https://doi.org/10.1016/j.jalz.2011.03.003

Squire, L. R. (1981). Two forms of human amnesia: An analysis of forgetting. *Journal of Neuroscience*, *1*(6), 635-640.

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

Stamate, A., Della Sala, S., Baddeley, A. D., & Logie, R. H. (2022). The effect of selective retrieval practice on forgetting rates in younger and older adults. Psychology and Aging, 37(4), 431–440. https://doi.org/10.1037/pag0000691

Taconnat, L., Bouazzaoui, B., Bouquet, C., Larigauderie, P., Witt, A., & Blaye, A. (2022). Cognitive mechanisms underlying free recall in episodic memory performance across the lifespan: testing the control/representation model. *Psychological Research*, 1-19. https://doi.org/10.1007/s00426-022-01736-1

Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica, 26, 24-36.

Trahan, D. E., & Larrabee, G. J. (1992). Effect of normal aging on rate of forgetting. Neuropsychology, 6(2), 115–122. https://doi.org/10.1037/0894-4105.6.2.115

Vale, F. A., Balieiro-Jr, A. P., & Silva-Filho, J. H. (2012). Memory complaint scale (MCS): Proposed tool for active systematic search. Dementia & neuropsychologia, 6, 212-218.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: sleep-dependent motor skill learning. *Neuron, 35*(1), 205–211. https://doi.org/10.1016/S0896-6273(02)00746-8

Walsh, C. M., Wilkins, S., Bettcher, B. M., Butler, C. R., Miller, B. L., & Kramer, J. H. (2014). Memory consolidation in aging and MCI after 1 week. *Neuropsychology*, *28*(2), 273. https://psycnet.apa.org/doi/10.1037/neu0000013

Wechsler, D. (1997). *Wechsler memory scale (WMS-III)* (Vol. 14). San Antonio, TX: Psychological corporation.

Weston, P. S., Nicholas, J. M., Henley, S. M., Liang, Y., Macpherson, K., Donnachie, E. & Zeman, A. Z. (2018). Accelerated long-term forgetting in presymptomatic autosomal dominant Alzheimer's disease: a cross-sectional study. *The Lancet Neurology*, *17*(2), 123-132. https://doi.org/10.1016/S1474-4422(17)30434-9

Wheeler, M. A. (2000). A comparison of forgetting rates in older and younger adults. Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition, 7(3), 179–193. https://doi.org/10.1076/1382-5585(200009)7:3;1-q;ft179

Wilson, B. A., Baddeley, A., Evans, J., & Shiel, A. (1994). Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological rehabilitation*, *4*(3), 307-326. https://psycnet.apa.org/doi/10.1080/09602019408401463

World Health Organization. (2020, Sept 21). *Dementia* https://www.who.int/news-room/fact-sheets/detail/dementia

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341

Zimmermann, J. F., & Butler, C. R. (2018). Accelerated long-term forgetting in asymptomatic APOE ε4 carriers. *The Lancet Neurology*, *17*(5), 394-395. https://doi.org/10.1016/S1474-4422(18)30078-4