University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

04 University of Plymouth Research Theses 01 Research Theses Main Collection

2021

Understanding the Contribution of Meaningful Processing to the Testing Effect

Hendry, Sarah

http://hdl.handle.net/10026.1/17780

http://dx.doi.org/10.24382/427
University of Plymouth

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognize that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.

UNIVERSITY OF
PLYMOUTH
Doctoral College

Understanding the Contribution of Meaningful Processing to the Testing Effect

by

Sarah Hendry

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology

July 2021

Acknowledgements

I would like to thank the School of Psychology at the University of Plymouth for awarding me the funding to carry out the work included in this thesis. Particular thanks go to my primary supervisor Dr Michael Verde who gave me opportunity to explore a diverse range of questions as part of this PhD, which has been rewarding. I would like to thank my secondary supervisor, Prof Tim Hollins, who helped with thinking outside the box.

It has been great to have support across the Psychology department more broadly, in particular the tech office team, wider staff and students. It has been a productive and vibrant environment to work in.

As a doctoral student, I have been able to access the training and support offered by the doctoral college. The training they offer and opportunities to showcase work has been extremely useful.

Thanks to Will Brigers for help with data collection.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. Work submitted for this research degree at Plymouth University has not formed part of any other degree either at Plymouth University or at another establishment. Relevant scientific seminars and conferences were regularly attended at which work was often presented.

Word count for the main body of this thesis: **57,479**

Signed: *Sarah Hendry*

Date: July 2021

Conference presentations:

Experimental Psychological Society UCL. Poster presentation: *Assessing the contribution of meaningfulness in study materials to the testing effect.* 2020.

Abstract

Sarah Hendry: Understanding the Contribution of Meaningful Processing to the Testing Effect

THE testing effect is an interesting phenomenon, with a wealth of support for its robustness. The basic idea is that attempting to learn something to retain over time is more fruitful when items are retrieved from memory or *tested*, than when items are restudied. This area of research has seen much attention in recent years, with the focus moving away from the conditions under which the effect can be found, to understanding the specific mechanisms that drive the effect. However, progress in this regard appears to be slow and contradictory. This thesis aims to address the gap in our understanding by exploring the concept of meaningful processing in relation to the testing effect. Here, how differences in meaningful processing relate to the testing effect is explored in text materials based on areas in the literature that have shown promise. More specifically meaningful processing will be explored herein based on; in chapters 2 and 3, how amenable study items are to meaningful, elaborate processing during retrieval (experiments 1-4), in chapter 3, whether there is a retrieval benefit associated with study items more meaningfully processed than less meaningfully processed, based on their structure (experiments 5 & 6). In chapter 4, differences in meaningful processing are further explored based on properties of the practice task as opposed to the study materials (experiments 7-10). Chapter 5 concludes that the results show little evidence that differences in meaningful processing of the study materials alter the magnitude of the testing effect (experiments 1-6), but some evidence that differences in meaningful processing during the practice task alter the magnitude of the testing effect (experiment 7).

Contents

Acknowledgements v

Author's declaration vii

Abstract ix

Contents xi

List of Figures xv

1 Findings and theories of the testing effect and exploring meaningful processing 1

1.1 Introduction . . . 1

1.2 The Testing Effect Paradigm . . . 6

1.3 Key Findings . . . 7

1.4 Theories of the Testing Effect . . . 13

1.4.1 Bifurcated Distribution Account . . . 13

1.4.2 Transfer Appropriate Processing . . . 14

1.4.3 Effortful Retrieval . . . 15

1.4.4 Episodic Context Account . . . 17

1.4.5 Elaboration Theory . . . 20

1.5 Focus on Meaningful Processing . . . 26

2 Revisiting the foundations of the Elaborate Retrieval Hypothesis 31

2.1 Introduction . . . 31

2.2 Experiment 1 . . . 35

2.2.1 Methods . . . 36

2.2.2 Results . . . 40

2.2.3 Discussion . . . 45
2.3 Experiment 2 . . . 47
2.3.1 Methods . . . 47
2.3.2 Results . . . 49
2.3.3 Discussion . . . 52
2.4 Experiment 3 . . . 55
2.4.1 Methods . . . 55
2.4.2 Results . . . 56
2.4.3 Discussion . . . 60
2.5 General Discussion: Experiments 1-3 . . . 61

3 Meaningful processing via mediation and structural coherence 65
3.1 Introduction . . . 65
3.2 Experiment 4 . . . 69
3.2.1 Methods . . . 71
3.2.2 Results . . . 76
3.2.3 Discussion . . . 80
3.3 Experiment 5 . . . 84
3.3.1 Methods . . . 85
3.3.2 Results . . . 88
3.3.3 Discussion . . . 90
3.4 Experiment 6 . . . 91
3.4.1 Methods . . . 92
3.4.2 Results . . . 95
3.4.3 Discussion . . . 97
3.5 General Discussion: Experiments 4-6 . . . 98

4 Meaningful processing during retrieval practice 105
4.1 Introduction . . . 105
4.2 Experiment 7 . . . 111

4.2.1 Methods . . . 112
4.2.2 Results . . . 116
4.2.3 Discussion . . . 118
4.3 Experiment 8 . . . 121
4.3.1 Method . . . 122
4.3.2 Results . . . 124
4.3.3 Discussion . . . 126
4.4 Experiment 9 . . . 128
4.4.1 Method . . . 129
4.4.2 Results . . . 130
4.4.3 Discussion . . . 134
4.5 Experiment 10 . . . 135
4.5.1 Methods . . . 136
4.5.2 Results . . . 139
4.5.3 Discussion . . . 141
4.6 General discussion: Experiments 7-10 . . . 143

5 Discussion **147**

5.1 Meaningful Processing Explored . . . 147
5.2 Summary of Results . . . 148
5.2.1 Chapter Two Results . . . 148
5.2.2 Chapter Three Results . . . 150
5.2.3 Chapter Four Results . . . 153
5.3 Null Findings . . . 156
5.4 Implications for Theory . . . 161
5.4.1 Elaborate Retrieval Hypothesis . . . 161
5.4.2 Mediator Effectiveness Hypothesis . . . 165
5.4.3 Constructed Retrieval . . . 167
5.4.4 Bifurcated Distribution . . . 167

5.4.5 Episodic Context Account 168
5.5 Implications for Practice 169
5.6 Future Work 171
5.7 Conclusion 173

List of references **175**

Appendices **189**

A Chapter Two Materials **189**

A.1 Word pairs for Exps 1, 2 & 3 190

B Chapter Three Materials **193**

B.1 Word pairs for Exp 4 194
B.2 Coherent Text Materials for Exp 5 200
B.3 Practice & Final Test Items Exp 5 203
B.4 Study Items & Test Questions Materials for Exp 6 206

C Chapter Four Materials **213**

C.1 Study Materials for Exp 7 214
C.2 Practice Test Materials for Exps 8 & 9 223
C.3 Study & Test Materials for Exp 10 227

List of Figures

2.1 Mean target retrieval as a function of test phase and association strength in experiment 1. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 43

2.2 Mean target retrieval at final test as a function of practice task and association strength in experiment 1. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 45

2.3 Mean target retrieval as a function of test phase and association strength in experiment 2. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 51

2.4 Mean target retrieval at final test and association strength as a function of practice task in experiment 2. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 53

2.5 Mean target retrieval as a function of test phase and association strength in experiment 3. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 59

2.6 Mean target retrieval at final test as a function of practice task and association strength in experiment 3. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 60

3.1 Mean target retrieval at final test as a function of practice task and definition language in experiment 4. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 81

3.2 Mean target retrieval at final test as a function of practice task and text structure in experiment 5. Error bars depict the standard error of the mean. 90

3.3 Mean target retrieval at final test as a function of practice task and text structure in experiment 6. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 98

4.1 Mean target retrieval at final test as a function of practice task and practice type in experiment 7. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 119

4.2 Mean target retrieval at final test as a function of practice task and practice type in experiment 8. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 127

4.3 Mean target retrieval at final test as a function of practice task and practice type in experiment 9. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 133

4.4 Mean target retrieval at final test as a function of practice task and practice type in experiment 10. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008). 141

5.1 Forest plot of mini-meta analysis results, weighted by sample size for Restudy versus Test Practice. *Denotes a significant testing effect **Denotes a significant negative testing effect (restudy advantage). Exp 4 compares restudy and test without feedback. 161

Chapter 1

Findings and theories of the testing effect and exploring meaningful processing

The testing effect literature reaches as far back as the late nineteenth century. Many studies since then have explored the conditions under which testing can be beneficial to long-term memory. However, theoretical progress in this area has stalled and new approaches are required to further our understanding. This chapter will outline the key findings in the literature as well as the key explanations for those findings. The chapter will highlight how meaningful processing is a useful area for exploration in tackling the current gap in our understanding and outline how this concept will be further explored within.

1.1 Introduction

In memory research, retrieval practice is attempting to retrieve previously studied information from memory, in order to avoid forgetting that information. The testing effect is the robust finding that testing, or *retrieval practice*, is a more useful learning tool than an alternative study activity for later memory retention. The most recent theoretical advances of the testing effect are now over five years old and the exploratory work since has shown limited gains in our understanding. Therefore, progress in understanding the testing effect phenomenon seems to have slowed. This chapter highlights one area where essential work into the testing effect has been all but overlooked, which is the contribution of meaningful processing to the testing effect, and furthermore details how I will attempt to bridge the gap in understanding in this regard.

The memory phenomenon known as the testing effect was borne out of early studies looking into how forgetting occurs and how forgetting could be mitigated through rehearsal (for example, Ebbinghaus, 1913; Gates, 1922 and later Allen, Mahler, & Estes, 1969; Bregman & Wiener, 1970). Findings suggested that repeatedly accessing a

memory through retrieval practice, could influence how retrievable an item was at a later point (Allen et al., 1969). In addition, retrieval practice was found to be more beneficial for subsequent retrieval success than restudying the information alone (Hogan & Kintsch, 1971).

Communication of the early success of retrieval practice for long-term memory has seen it adopted in widespread fashion in education for summative assessment (Adesope, Trevisan, & Sundararajan, 2017). Beyond this however, testing improves the accuracy of individuals' knowledge about their own learning (Dunlosky & Lipko, 2007). It also measurably improves outcomes in comparison to alternative study techniques that students are known to use, such as rereading from text books (Karpicke, Butler, & Roediger, 2009), which offer limited benefits to learning (Callender & McDaniel, 2009). More recently, low stakes quizzing, as opposed to exam testing, has also been shown to reduce individuals' test anxiety (Agarwal, D'Antonio, Roediger, McDermott, & McDaniel, 2014). With our increased understanding of the widespread benefits associated with testing, focus has moved to capitalising on this benefit and promoting testing for low stakes quizzing in everyday educational practice (Roediger & Karpicke, 2006a), as opposed to summative assessment alone. However, this poses issues for teachers by increasing the effort it takes to prepare the associated educational materials. In addition, testing is often a less preferred learning technique by students (Karpicke et al., 2009). Weighing up the evidence and practical issues of implementation is the role of policy makers in this area (Buck, Ritter, Jensen, & Rose, 2010) and any decision in this regard has widespread consequences. Therefore it is highly important that useful evidence continues to come to the fore in this domain, so that any decisions regarding application of this method can be made from a truly informed perspective.

Results that have come since Roediger and Karpicke's (2006a) recommendation convey a wealth of evidence for the benefit of testing in retaining specific information. Indeed, a recent meta-analysis found testing can be used to boost learning at all levels of education (Adesope et al., 2017). However, there are still many areas where there are more questions than answers. For example, there is still much debate as to

whether retrieval practice can be universally applied to increase recall for all forms of information being learned. Early work suggested that different information could benefit from testing to different extents (Bregman & Wiener, 1970; Gates, 1922). These ideas have frequently reappeared in the literature (Bouwmeester & Verkoeijen, 2011; de Jonge, Tabbers, & Rikers, 2015; Roelle & Berthold, 2017; Roelle & Nückles, 2019; Rowland, 2014; Schneider, Körkel, & Weinert, 1989; Van Gog & Sweller, 2015), yet to date there has been limited study devoted to exploring this concept in any detail.

In addition, there is a gap in understanding how changes to the retrieval practice task contribute to the testing effect. However, there is some evidence to suggest that differences in the retrieval task can influence the size of the testing effect. For example cued recall, whereby cues from previously studied items are given to aide recall, and free recall, whereby no prompt is given, yield larger testing effects, than recognition, whereby participants are required to answer whether they have seen the stimulus item previously or are given multiple options from which they are required to recognise the correct answer (Greving & Richter, 2018; Rowland, 2014). One explanation for this is that increased difficulty is inherent in cued recall and free recall in comparison to recognition tasks (E. L. Bjork, Bjork, et al., 2011). However, beyond the suggestion of increased difficulty leading to the testing effect benefit, understanding has been slow.

Furthermore, in relation to applied issues in this area, there has been less exploration of how testing might benefit broader knowledge transfer, such as applying knowledge to solve novel problems (Pan & Rickard, 2018; Van Gog & Sweller, 2015). Although some promising results have been found in this regard (McDaniel, Howard, & Einstein, 2009; Pan & Rickard, 2018), these effects are typically smaller (Pan & Rickard, 2018) than ordinary testing effects (Rowland, 2014). Therefore, while the signs suggest testing is an effective tool for educational practice, there are gaps in our understanding of the mechanisms behind these benefits.

I will attempt to address these gaps in our understanding through the common thread shared by this previous work. That is, that differences in the processing of the materials, either during the study task or the practice task, relate to changes in the

magnitude of the testing effect. These differences are considered herein as the degree of meaningful processing achieved during the particular task in focus.

Understanding the meaning of the materials being studied, for all students, is key to positive educational outcomes (deWinstanley & Bjork, 2002). Current theory and evidence of the testing effect, as already outlined, suggests meaningful processing is an important aspect of why testing is so beneficial. Yet, it has only been explored in superficial detail. Meaningful processing will be operationalised in different ways here to fully explore its contribution to the testing effect. By meaningful processing it is meant processing with *great value or significance*. Therefore I will explore how processing items in ways that change their value or significance impacts the magnitude of the testing effect. Key work, that has influenced the work in this thesis will be discussed further in this chapter, but falls broadly into three main areas.

The first concept of meaningful processing I explore here is the semantic relatedness of the study items. Studies have found that semantically related items are more memorable, less amenable to interference (Goodmon & Anderson, 2011) and based on their processing can benefit more from testing (Carpenter, 2009; Pyc & Rawson, 2010). For example, Goodmon and Anderson (2011) found that items that are semantically related to items retrieved are less likely to be forgotten when not tested. This is thought to be due to retrieval of one item simultaneously activating the overlapping semantic concept of the item not recalled. This suggests that semantic information activated during retrieval can be useful for subsequent retrieval of related information. Theoretical work relevant to this in the testing effect literature has suggested that semantic processing is central to testing effects (Carpenter, 2009; Pyc & Rawson, 2010). For example, Carpenter (2009) found that cue-target pairs that were more strongly associated semantically benefited more from testing than weakly associated pairs. Yet, still very few studies have explored this directly or in any detail. With much intuitive appeal, there is a tendency for this work to suffer less scrutiny. Therefore, it is important that work into semantic relations of the study items be fully explored in relation to the testing effect as it forms a very natural and potentially easy path to application.

For example, differences in the benefit of retrieval practice for semantically related and unrelated items would result in differences in the practical guidance for studying each set of items. Following this future research could look to further identify the optimal conditions for studying each set.

The second concept of meaningful processing I explore here is how well study items can be integrated as a whole. In particular, to what extent a disruption to meaningful processing can impact the benefit associated with retrieval practice. Studies have shown that items that are organised in a more coherent structure tend to benefit more from retrieval practice (Rowland, 2014). These materials are less likely to be forgotten when only some of the materials are retireved (M. C. Anderson & McCulloch, 1999; Chan, 2009), suggesting that these materials are processed more completely or more meaningfully. While more recent work further reiterates that highly coherent information is processed differently (de Jonge et al., 2015; Hostetter, Penix, Norman, Batsell Jr, & Carr, 2019), this concept has not been directly examined under typical testing effect conditions, namely, within the same experiment and with an appropriate restudy control condition that matches both the time spent on task and the number of times items are studied. Therefore, because work that has shown promising results in the retrieval-induced forgetting literature (Chan, 2009), a related field to the testing effect, did not use an appropriate restudy control, further enquiry is required to demonstrate the application of these results to the testing effect.

The third and final concept of meaningful processing I explore here relates to the retrieval practice task. Some work on the benefits associated with this comes from comparisons between different retrieval practice tasks, for example short answer questions compared to multiple choice questions (Greving & Richter, 2018). Here learning via short answer questions was compared to multiple choice questions (without feedback) at time points of 1 week, 10 weeks or 23 weeks after a semester of learned content in a real-life educational context. Results revealed the short answer questions to be more beneficial to long-term retention than multiple-choice questions, but only for the items that were retrieved well to begin with on average. Results were suggested

to be in line with an effortful processing account of testing effects, whereby increased effort at the time of retrieval in this case retrieving via short answer as opposed to multiple choice questions leads to a testing advantage over restudy.

Further work relating to the contribution made by the retrieval task to the testing effect comes from the transfer testing effect literature (Hinze, Wiley, & Pellegrino, 2013; Pan & Rickard, 2018), whereby more elaborate processing during retrieval practice is shown to benefit conceptual knowledge transfer. For example, Hinze et al. found that when participants organised knowledge into their own words or for comprehension during retrieval practice, there was a greater transfer benefit over restudy. The benefit occurred on inference questions at final test, in comparison to when instructions were to attempt to recall items in their original form. However, once more this work has not demonstrated these benefits under matched control conditions, limiting the scope for understanding, interpretation and application.

This chapter will now outline the key findings relating to the testing effect and further explore the theoretical explanations for these key findings. It will conclude by detailing how meaningful processing will be explored across the ten experiments enclosed.

1.2 The Testing Effect Paradigm

Testing effect research came from early work on how to preserve memory. In realising the utility of this approach, it has been necessary to compare retrieval practice to alternative learning strategies, such as concept mapping (Karpicke & Blunt, 2011), self-referential elaboration (Endres, Carpenter, Martin, & Renkl, 2017), as well as keyword mnemonic techniques (Karpicke & Smith, 2012) and note-taking (Rummer, Schweppe, Gerst, & Wagner, 2017). These comparisons make for a diverse and necessary literature for understanding how best to apply the testing effect to an educational context. However, a more constrained standard has also been established to be able to examine the mechanisms behind the testing effect and more easily compare results across studies.

The typical testing effect is a comparison between a retrieval practice test and a

repeat study, or *restudy*, opportunity. The benefits of comparing testing to a restudy control are two-fold. Firstly, it ensures that retrieval practice benefits cannot be explained by the amount of time reexposed to the study materials during the practice task alone, as restudy groups spend the same amount of time restudying the materials as retrieval groups spend attempting to retrieve the materials. Secondly, having a restudy control makes it possible to compare retrieval practice to a study strategy that students often employ and find less taxing than retrieval practice (Karpicke et al., 2009).

The testing effect paradigm usually involves three phases. The first is a study phase, in which all participants engage with the learning materials in the same way, typically by reading through them or passively studying them. This is followed by a practice phase, or *learning phase*, in which an experimental manipulation of some form occurs. Usually, participants are either tested on the learning materials (test practice group) or restudy the same learning materials again (restudy practice group), with the time on each task being matched. Rereading or restudying is the act of reading or studying without being engaged with an additional task like note-taking. Finally, participants complete the final test phase, where all participants take the same criterion test on the learning materials. The final test is typically given during the same experimental session or at delay of up to several days. Research compiled over the last several decades has resulted in some common findings that are outlined below.

1.3 Key Findings

The key findings from the testing effect literature were neatly summarised in a meta-analysis conducted by Rowland (2014). Rowland's meta-analysis included only testing effect studies where the timing for the restudy and test practice conditions were matched, which included 159 effect sizes from 61 studies. The analysis assessed to what extent the different design characteristics of the studies contributed to the testing effect. The design characteristics assessed included whether feedback was given or not. When feedback is given it often involves presenting the intact study item that the

participant had attempted to retrieve immediately prior, for several seconds. The analysis also included whether the design utilised mixed lists or pure lists, which is whether the practice task for each participant contained only study items or retrieval practice items (pure lists) or a mix of both items (mixed lists). The analysis also assessed the contribution of a design that was within or between-subjects, so whether participants completed both a retrieval practice task and a restudy task (within-subjects) in the study or only completed one or the other (between-subjects).

In addition, the analysis assessed the period of delay between the initial and final test, which was in the order of minutes to days. The analysis also looked at the contribution of the format of the initial and final tests. For example, whether the initial retrieval practice task was a recognition test, a multiple choice test, a cued recall test or a free recall test, with the same categorisation for the final test also. Finally, the analysis also assessed whether different properties of the stimuli contributed to the testing effect. For example, to what extent the testing effect is influenced by items' semantic organisation (semantically organised or not) and presentation format (lists of words or passages of text). The analysis was run on the total set of effect sizes and a subset of effect sizes, in which the retrieval practice task included a high level of re-exposure to the study materials. Here, a high level of re-exposure meant study materials that were retrieved with high accuracy during the retrieval practice task ($> 75\%$), or experimental designs that allowed participants to view the study materials again after failed retrieval or a retrieval attempt, through feedback. Several clear findings were reported.

The overall mean effect size for the testing effect across all studies analysed was estimated to be a medium effect size (Sawilowsky, 2009), g = 0.50 (CI [0.42, 0.58]). This is consistent with a less constrained more recent meta-analysis of the testing effect (Adesope et al., 2017), that included studies with a broader range of comparison conditions, for example alternative revision tasks such as concept mapping and no study filler tasks. For Rowland's subset of high exposure data, the size of the effect was slightly higher, g = 0.66 (CI [0.56, 0.75]). Although the inclusion of feedback is likely responsible for the size of the difference as studies with feedback resulted

in a significantly larger testing effect (g = 0.73, CI [0.61, 0.86]) than studies without feedback (g = 0.39, CI [0.29, 0.49]).

For studies without feedback the size of the testing effect increased with the likelihood of retrieval during the initial retrieval practice task. When the initial test performance was less than or equal to 50% no testing effect was found (g = 0.03, CI [-0.21, 0.27]). For studies where the initial test performance was between 51% and 75% a reliable testing effect was found (g = 0.29, CI [0.09, 0.49]). Studies in which initial retrieval was above 75% demonstrated the largest benefit of testing (g = 0.56, CI [0.42, 0.70]). These results could help us to understand the contribution of the restudy opportunity, which might be more beneficial in conditions where initial test accuracy is low (Van Gog & Sweller, 2015). Furthermore, as the impact of feedback is found to depreciate with each successive feedback opportunity and a benefit is not always seen (Adesope et al., 2017), it is necessary to also consider the direct effects of testing when feedback is not present (Karpicke, Lehman, & Aue, 2014). Previous work (Kang, McDermott, & Roediger, 2007) has shown that feedback can boost performance associated with a retrieval task, which is thought to occur due to enhanced encoding strategies following incorrect response reveal, in addition to increased memory strength for items that are retrieved successfully. To avoid introducing mediating effects associated with the provision of feedback during retrieval practice, and the added possibility for feedback to interact with the meaningful processing manipulations of interest here, the studies herein will largely not include feedback in the design. In light of this the remaining summary of the key findings will only be based on the full data set, rather than the subset of high exposure data.

For the different design components there was a smaller effect size for within-subjects designs, whereby the restudy and retrieval practice task were completed by each participant (g= 0.43, CI [0.35, 0.52]), than for between-subjects designs, whereby each participant completed only one practice task (g= 0.69, CI [0.48, 0.89]). There was no difference as to whether the practice items were shown in a mixed list (g= 0.49, CI [0.37, 0.62]) or pure list format (g= 0.46, CI [0.34, 0.57]). These results suggest that the

testing effect is likely to be maximised with a between-subjects design for the restudy and retrieval practice task components.

There was also a clear finding for the presence of a test-delay interaction. Whereby, studies in which the delay between the initial test phase and final test phase was greater than one day produced larger testing effects (g= 0.69, CI [0.56,0.81]) than studies in which the delay was less than one day (g= 0.41, CI [0.31, 0.51]). However at both retention periods a reliable testing effect was found, consistent with the results found by Adesope et al. (2017), suggesting that testing positively impacts the forgetting rate of the items studied. More widely some studies have found that short retention periods do not always result in a testing effect. One study that exemplifies this finding is by Roediger and Karpicke (2006b), who varied the length of the delay to the final test. They found that when the delay to the final test was five minutes, there was a benefit of restudy over testing. However, when the delay to the final test was 2 days, the pattern had reversed and there was now a large benefit for testing (Sawilowsky, 2009), d= 0.95. The finding of a restudy benefit with an immediate delay to final test is known as a negative testing effect and is thought to be due to the restudy condition being reexposed to all items, while the test condition (in the absence of feedback) is only reexposed to the retrieved items (Roediger & Karpicke, 2006b; Wheeler, Ewers, & Buonanno, 2003). Whereas, the reverse pattern at delay is thought to reflect a reduced rate of forgetting for retrieved items, which the restudy items do not benefit from. In this study it is interesting to note that at the final tested time point of one week this difference was not any larger, d= 0.83, which is consistent with the results of Adesope et al. (2017). These findings highlight that although a delay is beneficial, the nature of this benefit is not likely to be linear in function and substantial delays are likely to show limited protection against forgetting (Chan, 2010). Results therefore show that a delay in the order of a few days is likely to be an optimal retention period to utilise.

One clear finding relevant to the focus of meaningful processing, that emerged from the meta-analysis is the test format of the initial task and the final task. Results showed that practice tasks that employed a cued recall retrieval practice task yielded larger

testing effects (g = 0.61, CI [0.52, 0.69]) than those that employed free recall (g = 0.29, CI [0.07, 0.52]) or recognition practice tests (g = 0.29, CI [0.10, 0.47]). This suggests that difficulty in the practice task, associated with having fewer or no cues during the retrieval practice task (free recall), is not alone responsible for the benefits associated with retrieval practice. Here however, there is evidence that aspects of the practice task are important for the magnitude of the testing effect. Results for the final test format showed similar results with cued recall (g = 0.57, CI [0.46, 0.68]) and free recall (g = 0.49, CI [0.34, 0.63]) final tests giving larger testing effects than recognition tests (g = 0.31, CI [0.15, 0.46]). These results indicate that initial tests are most beneficial when in a cued recall format and final tests are equally beneficial in either a free recall or cued recall format.

Rowland's meta-analysis also examined the influence of moderators relevant to meaningful processing of the study materials. This was looked at in three different ways, based on the different properties of the stimuli. Firstly, stimulus type, as examined by; whether the study items were lists of individual words, related cue-target pairs or prose passages. The results suggested that the organisation of the stimulus type moderated the testing effect, whereby more organised materials resulted in larger testing effects, for example the prose (g= 0.58, CI [0.34, 0.82]) and cue-target pairs (g= 0.59, CI [0.49, 0.70]), compared with the less organised lists of words (g= 0.39, CI [0.24, 0.53]) and study items that did not fit into a category (g= 0.27, CI [0.06, 0.48]). This is consistent with the results found by Adesope et al. (2017), whereby passage learning resulted in larger testing effects (g= 0.71) than the learning of word lists (g= 0.56). Here we have our first evidence that meaningful processing relating to the structural properties of the materials influences the testing effect.

The second moderator in Rowland's meta-analysis relevant to meaningful processing of the study materials was the relationship between the cue and target where the stimuli was made up of word pairs. This moderator was made up of five different levels; same (recognition), non-semantic, semantic unrelated, semantic related and none (free recall). There was no heterogeneity between the levels, although the results did

suggest numerically that more meaningfully related items, by way of semantically related pairs (g= 0.66, CI [0.51, 0.82]) benefited more from testing than non-semantically related cue-target pairs (g= 0.54, CI [0.42, 0.66]). Furthermore, findings from studies that have manipulated the relatedness of the study materials within the same experiment suggest that this is an important factor to the magnitude of the testing effect (Carpenter, 2009; Carpenter & Yeung, 2017; Pyc & Rawson, 2009). Therefore, while the meta-analysis does not give a strong indication that semantic relationship between cue and target pairs are important for the testing effect, more work is required to fully understand this concept.

The final moderator relevant to meaningful processing in the study materials was based on how the stimulus interrelations of the materials influenced the testing effect. This had four levels; prose, categorical, no relation, other. The meta-analysis found no heterogeneity in the effect sizes between these levels, suggesting that the stimulus interrelations in the study materials might not be important for the mechanisms of the testing effect. However, this result reflects properties manipulated between studies, not within studies, which typically utilise different design elements. Therefore, the result conveys a somewhat crude application of meaningful categorisation that may or may not be relevant to the focus of meaningful processing herein. While the benefit of the meta-analysis is that it is a comparison between studies, the different experimental design features that it encompasses have been shown to influence the magnitude of the testing effect (Mulligan, Susser, & Smith, 2016) and could be masking the more nuanced factors being assessed in this moderator analysis. As noted by Karpicke (2017), how materials are manipulated within an experiment has been scarcely explored in relation to the testing effect and could be useful for theoretical developments in relation to the testing effect. Elsewhere in the related literature of retrieval-induced forgetting, associations between study items has been looked at in more detail and results suggest semantic categorisations can influence retrieval processes (M. C. Anderson & McCulloch, 1999).

Therefore, while there is a clear finding across two different meta-analyses that

some structural components are relevant to the magnitude of the testing effect, direct comparisons within the same experiment are scarce in the testing effect literature, particularly under the conditions that are optimal for studying the direct effects of testing and the conditions we now understand are likely to boost the testing effect.

One thing to note with Rowland's meta-analysis is that the studies included were heavily weighted on the results of paired associates and lists of single words which made up over 80% of the sample, rather than longer text study materials which made up less than 15% of the sample. Therefore, how prescriptive these results are for more educationally typical study materials is difficult to say. However, as the meta-analysis focused on studies that contained a restudy matched time control task and this is also a feature of the studies in this manuscript, it was necessary to outline a summary of relevant findings here. Now that I have outlined the main findings associated with the testing effect, I will move on to highlighting the key explanations for these findings.

1.4 Theories of the Testing Effect: Explaining the Key Findings

The key findings outlined above involve the inclusion of feedback and initial accuracy levels, the test-delay interaction, the nature of the retrieval tasks and the nature of the study materials. Below I will outline how these different results have been explained.

1.4.1 Bifurcated Distribution Account

The bifurcated distribution account offers a useful description of how retrieval practice produces higher accuracy levels, the test-delay interaction and a benefit associated with feedback (Kornell, Bjork, & Garcia, 2011). This account suggests that information that is successfully retrieved, due to lying above a memory strength threshold at the time of the initial test, gains additional memory strength that allows for easier recall at a later point. This is because the memory strength boost offered by successful retrieval is more potent than the memory strength boost offered by a restudy opportunity. Therefore, when initial accuracy is high, or feedback is present a test-delay interaction can be seen, because items that are retrieved accurately receive a boost in memory strength through retrieval. In the case of feedback this will include boosting the

items not retrieved in the same way as restudied items, while preserving the retrieval boost to items accurately retrieved. As the time to the final test increases, the discrepancy widens between the memory strength associated with the retrieved items and the restudied items. This accounts for the test-delay interaction typically seen in testing effect studies (Rowland, 2014), because more items in the retrieved set will lie above the memory strength threshold at the final test than those in the restudy set. While there is key evidence that these are all features of the benefit of retrieval practice, this account does not provide an explanation of the mechanisms of retrieval practice that offer this boost.

1.4.2 Transfer Appropriate Processing

Early memory research led to ideas that aspects of the learning environment, or context, provide useful cues during subsequent retrieval. These ideas emphasised that successful storage of encoded information was vital for successful retrieval (Tulving & Thomson, 1973). For example, early work examined how changes to the context of the study and retrieval environments influenced memory, revealing memory benefits associated with reinstating the original learning context during a retrieval attempt (Godden & Baddeley, 1975). However, results are not limited to the physical context, simply remembering the context is enough to demonstrate superior performance (S. M. Smith & Vela, 2001). The transfer appropriate processing account suggests that the practice test allows students to mentally experience the final testing parameters and it is this similarity between the mental processing required at each point that confers an advantage for retrieval practice (Morris, Bransford, & Franks, 1977).

In terms of the testing effect therefore, the transfer appropriate processing account suggests memory success relies on a match in cognitive processing between practice and final test. Rowland's meta-analysis did not find clear support for the transfer appropriate processing account, due to the fact that matched initial and final test formats did not show larger testing effects than mismatched tests, however, it is worth making clear that some recent findings still find support for this account (Adesope et al., 2017).

In addition, current perspectives share common ground with this approach (Karpicke et al., 2014), therefore it is useful to outline the explanatory power of this outlook.

Evidence can still be found in favour of this approach when the task is more challenging or applied (Hinze et al., 2013; Larsen, Butler, Lawson, & Roediger, 2013). Larsen et al. (2013) for example, in an applied examination of the utility of retrieval practice, compared different retrieval practice techniques on the retention of learned clinical procedures. Results showed that practice in a given technique transferred best to final exam performance in that technique. One criticism of the transfer appropriate account as an explanation of retrieval practice success, is that we do not learn anything about the underlying processes involved (Bradshaw & Anderson, 1982) or the conditions under which we are likely to find a testing effect. And due to this lack of specificity, many studies are seen to be at odds with this perspective (Carpenter & DeLosh, 2006; Rowland, 2014; S. M. Smith, Glenberg, & Bjork, 1978). Examinations of this concept have given rise to ideas that have attempted to address the lack of specificity, such as the elaborate retrieval hypothesis (Carpenter & DeLosh, 2006) and the episodic context account (Karpicke et al., 2014), which will be detailed below.

1.4.3 Effortful Retrieval

One finding that has been persistent in the testing effect literature is the idea that retrieval practice represents more effortful processing than restudy or rehearsal practice and that somehow this is leading to a memorial benefit for the retrieved items. The additional effort or difficulty comes from needing to locate the memory for previously presented information during retrieval, which is not required during restudy. It is this desirable difficulty, or added difficulty that results in a benefit, which makes the testing experience beneficial (E. L. Bjork et al., 2011). Results from an early study in this area made more specific predictions based on findings in the free recall of word lists (Craik, 1970). Craik found that items presented most recently in the list were often retrieved immediately and more successfully, in comparison to items presented earlier in the word list. Items presented earlier in the list tended to be recalled later in the free recall

sequence and with less frequency. However, upon a second free recall attempt of the list, the items that were retrieved from the earlier portion of the list, but later in the recall sequence tended to be recalled to a greater extent than the items that were recalled from the end of the presented list and earlier in the recall sequence. This led Craik to suggest that increased difficulty at the time of retrieval, which at an item level meant more difficulty retrieving items that had not been presented recently in the list, seemed to increase the success associated with retrieving that item at a later date.

Many subsequent studies have found evidence consistent with the idea of increased effort or difficulty during the retrieval practice task being beneficial for subsequent retrieval success. This difficulty has taken many forms, such as fewer cues being present during the initial retrieval attempt leading to a memorial benefit (Carpenter & DeLosh, 2006; Kang et al., 2007; Rowland, 2014). In addition, longer delays between the initial retrieval and the final retrieval attempt are also thought to represent increased difficulty, that results in a benefit (Karpicke et al., 2014). Furthermore, expanded retrieval, whereby retrieval attempts are separated in time, as opposed to massed and repeated immediately, tend to also increase the benefit of retrieval (Karpicke & Roediger, 2007). In a similar way, when items are repeatedly retrieved at longer lags between items as opposed to shorter lags between items, longer lag repeated testing is more beneficial to subsequent retrieval (Rawson, Vaughn, & Carpenter, 2015). In line with these findings, self-report methods tend to show testing conditions as being experienced as more difficult than restudy conditions and result in lower confidence in performance (R. A. Bjork, Dunlosky, & Kornell, 2013). Whilst this looks to be compelling evidence for the effortful processing explanation of the testing effect, the conditions under which effortful processing could occur appear to be endless. Furthermore, this approach undermines the importance of encoding and memory strength to subsequent retrieval (Craik, 2002; Kornell et al., 2011; Tulving & Thomson, 1973) and suffers from a lack of predictability over which difficulties will be desirable. This brings up a circular argument, as difficulties are only desirable when they result in an advantage over restudy. A recent review looked specifically at whether the testing effect could be achieved in

study materials that are particularly effortful to learn.

For example, Van Gog and Sweller (2015) suggested that the testing effect had not been well explored in relation to complex study materials. They reviewed studies that featured complex study information and found that this information did not seem to benefit from retrieval practice in the same way, often showing no testing effect or a negative testing effect. Complex information was defined as information containing components with high elemental interactivity. This translates into component parts that need to be integrated to be utilised. For example, understanding how electrical circuits work, requires understanding of the different component parts in order to apply this knowledge to novel problem solving. These materials arguably require a level of difficulty to learn, either during encoding or retrieval or both. It could be that this complexly related information requires increased levels of comprehension, or prior knowledge, for effective retrieval. Therefore, while difficulties in many forms do appear to increase the benefit of retrieval, this perspective is not very prescriptive and seems to have boundaries to its effectiveness (Karpicke & Aue, 2015; Van Gog & Sweller, 2015), which are still poorly understood.

Further to the explanatory accounts and broader memory theories of retrieval given above, there has been some attempt to theorise about the suggested mechanisms at play during retrieval practice that lead to a benefit over restudy practice more specifically. The more specific theoretical accounts of the testing effect are the episodic context account and elaboration theory under two guises, the elaborate retrieval hypothesis and the mediator effectiveness hypothesis. These accounts will be outlined below, before turning to the direct focus of this thesis.

1.4.4 Episodic Context Account

The episodic context account (ECA) highlights that the surrounding episodic context in which items are retrieved is important to the testing effect (Karpicke et al., 2014). One of the central ideas of the ECA is that in order to recall an item from memory, there is an active attempt during retrieval practice to reinstate the previous context that an item

was presented in. This account extends explanations of spacing effects, that suggest spacing effects are attributed to a combination of study-phase retrieval and contextual variation (Benjamin & Tullis, 2010). Spacing effects describe when spaces between study opportunities lead to better memory than massed study, or no spacing between presentations. Such explanations of spacing effects suggest that a degree of retrieval occurs when an item is re-presented. However, the ECA suggests that these effects reflect incidental retrieval that occurs during a restudy opportunity. Whereas, retrieval practice reflects intentional retrieval and therefore is likely to magnify the use of contextual features, which change as a result of increased temporal changes with spacing. Here, incidental retrieval is taken to mean any retrieval that occurs during restudy as a function of the repeated episode. Whereas, intentional retrieval is associated with a deliberate attempt to think back and remember the previous study episode. Some retrieval may occur during incidental retrieval (during restudy), but this occurs to a much lesser extent than during intentional retrieval (during retrieval practice).

In a similar way to the temporal context model of memory (Howard & Kahana, 2002), the episodic context account (ECA) relies on a constantly changing temporal context that is able to guide retrieval. Karpicke et al. (2014) proposed that the information retrieved during a practice test is bound to the features of the original presentation context, which then serve as retrieval cues at the final test. When retrieval of a previously presented item is attempted, features of the original presentation context are activated and bound to features of the current retrieval context. It is these updated contextual features that are better able to guide subsequent retrieval more efficiently than restudy. Context reinstatement is thought to occur during restudy although to a lesser extent than during retrieval, as less effortful reinstatement of the previous context is required. Support for this account is shown from studies that measure some form of temporal processing, for example the clustering of the retrieval order or an ability to temporally sort the previously presented items.

In this way, Whiffen and Karpicke (2017) found support for the episodic context account with a list discrimination task. After learning two lists of words, both restudy

and retrieval condition participants were presented the words again. In the retrieval condition participants were instructed to think back to which list the item was previously presented in. The retrieval group had both improved temporal clustering and retrieval rates compared to the restudy condition. This suggests that enhanced performance associated with retrieval practice could be due to context reinstatement. Similar results are reported when retrieval practice is compared to non-retrieval elaboration strategies (Lehman, Smith, & Karpicke, 2014), whereby a non-retrieval elaboration task results in inferior temporal memory than retrieval practice. Here, the argument is that elaboration accounts of retrieval practice do not explain the retrieval memory benefit as well as the context reinstatement account.

Whilst the changes in the temporal context can provide a constant way to apply the benefit of retrieval to many different types of stimuli, the account also suggests that when additional effort is required at a number of different stages in the retrieval process, this will prove to be more beneficial to a later retrieval attempt. For example, the ECA account depends on a changing temporal context to make sense of how retrieval practice enables more efficient later retrieval, suggesting that at a greater delay, or increased spacing to contextual reinstatement a greater testing benefit will result. This leads to the prediction that items benefit differently based on how difficult they are to retrieve.

In line with this, studies have shown that difficulty associated with temporal features of the study phase and retrieval practice phase do result in an increased testing effect. For example, Rawson et al. (2015), found that testing gains over restudy are greater when the items are presented with longer lags (35 items) between repeat presentations than at shorter lags (8 items). This result is more pronounced with a larger delay to final test (Carpenter & Yeung, 2017), and occurs when feedback is included (Pyc & Rawson, 2009). It does therefore seem, that timing mechanisms alongside increased difficulty in timing features are relevant to retrieval processes. However, there is a caveat in this theory that suggests that items will reach a threshold after repeated retrieval, in such that they become decontextualised and no longer require features of the episodic con-

text to be retrieved. This idea is supported by the notion that with repeated retrieval, the benefit of retrieval practice follows a negatively accelerated curve (R. A. Bjork, 1999). In addition, Pyc and Rawson (2007) found that longer lags between item presentations (23 vs. 5) are more beneficial for memory of the item when using drop out schedules, whereby to be learned items are dropped after one successful retrieval. This shows that even when all initial retrieval is successful, it is more beneficial at a later date when a larger delay or more difficulty to context reinstatement has occurred.

Some previous work however is not entirely consistent with the ECA, For example, Brewer, Marsh, Meeks, Clark-Foos, and Hicks (2010) found that when gender information of the previous presentation was retrieved during the practice task, then temporal information was not as accurate as when retrieval practice involved a list discrimination task. This suggests that the nature of the practice task influences what information is available on the final test. Other more recent work suggests that context updating might not just be limited to temporal features (Schwoebel, Depperman, & Scott, 2018) and memory for broader contextual features are enhanced by retrieval practice (Akan, Stanley, & Benjamin, 2018). An additional issue, that is common for all theoretical concepts in this area, is the issue of circularity (Karpicke, 2017), whereby, evidence of causation is taken from phenomenological outcomes. In other words, the fact that contextual features are boosted at the time of the final test, does not necessitate that processing of contextual features during retrieval practice was the source of this benefit. This will be an issue for future work to wrestle with.

1.4.5 Elaboration Theory

Ideas about the usefulness of elaboration to memory utilises the levels of processing account of memory, whereby depth of processing, indicative of semantic depth of processing, evolved to incorporate not only semantic depth of processing but also the spread of processing or how elaborately processed items are (Craik & Tulving, 1975; Moscovitch & Craik, 1976). This also gave rise to ideas about material appropriate processing, whereby the processing task is most useful when it reflects how the infor-

mation is likely to be used. For example, if you need to remember the colour a word is presented in, it is not helpful to concentrate on its semantic properties. Similarly, if you need to later remember the list membership of the item, it is not useful to practice remembering which gendered voice the item was presented in (Brewer et al., 2010).

Early evidence supports the idea that more relevant processing promotes recollection of this relevant information at a later point (Johnson-Laird, Gibbs, & De Mowbray, 1978). Further work suggested that processing relational elements of the studied items is particularly useful more broadly for later retrieval (Einstein, McDaniel, Owen, & Cote, 1990). Further to this Willoughby, Wood, and Khan (1994) found that elaborately interrogating information, by asking participants why questions during encoding is useful for memory retention, possibly due to its assistance with organising knowledge. Applied to the testing effect, retrieval practice is proposed to be beneficial because information, through more extensive processing channels, becomes integrated into long-term memory, as opposed to accessibility being temporarily boosted due to restudy. It is the additional processing that serves as cues to the items to be retrieved at a later time. Yet, the exact nature of what is helpful about the additional processing has not been clearly established.

In 2006 Carpenter and DeLosh suggested evidence for elaborative retrieval as an explanation of the testing effect. The study was designed to look at whether transfer appropriate processing or elaborate processing is more instrumental to the testing effect. Participants were instructed to retrieve items of single words from a previously presented list with as few cue letters as possible, although they could ask for as many letters as required to retrieve each word (experiment two). When fewer cues were used to retrieve the item, items were retrieved to a greater extent during a free recall task than the items retrieved with more cue letters initially. This finding was replicated in experiment three when the number of cues available during retrieval practice was directly manipulated. Authors suggested this to be evidence of more elaborate processing during retrieval leading to a greater benefit of retrieval practice. Meaning that when fewer cues are available during the retrieval practice task, the amount of elab-

orate processing required to retrieve the item increases. This elaborated information becomes associated with the item, which boosts the available cues during the final retrieval attempt (McDaniel & Masson, 1985). This idea was later specified to a greater extent by Carpenter (2009).

Elaborate Retrieval Hypothesis

Following on from the earlier study (Carpenter & DeLosh, 2006), Carpenter (2009) looked to formalise a new explanation of the testing effect drawing on the strengths of evidence from effortful processing and aiming to avoid the shortcomings associated with the transfer appropriate processing account. Carpenter sought to examine whether the elaborate retrieval hypothesis could explain the testing effect. The hypothesis proposed that retrieval practice benefitted more from elaboration than restudy practice. Elaboration occurs when searching memory for the correct item during retrieval practice, the items searched in memory become useful cues during the subsequent search at final test. Strongly and weakly associated word pairs were utilised as study materials across two experiments. Carpenter hypothesised that support for the elaborate retrieval hypothesis would be demonstrated if there was a larger testing effect seen for the weaker associate pairs than for the stronger associate pairs. This is because during the retrieval practice task it is suggested that the weaker associate pairs undergo a larger, more elaborate search of memory. This activates more items in memory that later serve as efficient cues to retrieval during the final test. This is the first testable way to assess an aspect of difficulty inherent in the study materials in relation to the testing effect.

The results of experiment one (Carpenter, 2009), where short mixed lists of strong and weak associated word pairs were learned (8 item lists) in a fully within-subjects design, showed a clear benefit for learning the weaker associate pairs through retrieval practice. This was revealed through greater retention rates of the weaker associates compared to the stronger associates from initial to final test, as indicated by a significant interaction. In addition, the final test performance comparison between re-

trieval practice and restudy, revealed an interaction in favour of better recall for retrieved weaker associates over restudy, than for the equivalent stronger associates. However, the second experiment which was fully between-subjects and where longer pure lists of weak and strong associates were learned (48 item list), did not show the same level of convincing evidence, as no testing effect advantage for the weaker associate was seen at final test based on absolute accuracy comparison. Yet, Carpenter reported that the weaker associates in both experiments demonstrated evidence of shallower forgetting curves. This was evidenced by the interactions between strong and weak associates from initial to final test in both experiments. There was also an inclusion of a conditional analyses for the testing effect data for experiment two, whereby final test accuracy for the retrieval practice group was calculated as a proportion of correctly recalled items during the initial test. In this analysis a weaker associate advantage was revealed. While conditional and absolute accuracy at final test have been used previously to calculate the testing effect, it is largely understood that a conditional analysis will exacerbate the role of item effects or memory strength in the testing effect (Kornell et al., 2011). There are additional questions surrounding the fact that the design in experiment two was between-subjects, which we now understand to evoke larger testing effects (Rowland, 2014), and included a longer list format, which we would hope the effect would be more robustly applied to. Taken together, this suggests that there could indeed be other explanations for these results, or at least these results should be followed up in relation to the phenomena we expect to find associated with the testing effect today.

The evidence reported by Carpenter appears to be compelling for an elaborate retrieval explanation of the testing effect. Whereby, retrieval practice shows a favourable boost to items that require more elaboration to retrieve (weaker associated items). Furthermore, this explanation benefits from having some intuitive appeal. However, criticisms of the ERH have been found more recently, where studies have failed to show that increasing elaboration during study practice is as beneficial as retrieval (Karpicke & Smith, 2012; Lehman & Karpicke, 2016). Evidently, there are some issues associated

with the ERH, however, the idea that elaboration of some kind is important for retrieval is still a supported notion in work on the testing effect. To this end an alternative iteration of the importance of elaboration to the testing effect has been presented and has received more empirical support, which is the mediator effectiveness hypothesis (Carpenter, 2011; Pyc & Rawson, 2010).

Mediator Effectiveness Hypothesis

The mediator effectiveness hypothesis, suggests that when mediating information is processed during retrieval, this can subsequently be used to effectively cue retrieval at a later date. Mediating information is thought of as information that can help form links between cues and targets. For example, mediating information might take the form of an associated semantic network being activated during the encoding phase, which subsequently assists retrieval in a similar way to the ideas contained in the elaborate retrieval hypothesis. However, inconsistent with the ERH, this perspective suggests that when stronger linking information is more accessible during retrieval practice, it is likely to be more beneficial to retrieval practice than weaker linking information.

Early work into the concept that supplementary or mediating information is beneficial to retrieval, found that implicit associations that are activated during stimulus list presentation could explain subsequent false recognition (Underwood, 1965). Results of this study showed that lures in a recognition test were falsely recognised based on the frequency of presentation of their associated target word, suggesting stronger activation of this information influenced subsequent memory for the item. Similar work has shown that implicit associations or mediating information can even be falsely recalled in a free recall task (Roediger & McDermott, 1995), suggesting that this information could be important for retrieval processes. Mediating information can also be supplied rather than implicit associations (Carpenter & Yeung, 2017), as when items low in meaning are supplied with meaningful contextual information, they become more memorable (Crouse, 1967). This concept is also relevant to work on learning new vocabulary, whereby keywords generated by the individual serve to form a helpful link between

familiar and unfamiliar concepts and guide retrieval (McDaniel & Pressley, 1989).

This idea was picked up and extended in relation to the testing effect in a study by Pyc and Rawson (2010). In this study participants were required to learn Swahili-English cue-target word pairs. While studying the pairs, participants initially generated mediating information to help them associate the cue with the target. The mediating information was to help them subsequently recall the pair and took the form of a word that sounded like the cue but was semantically similar to the target. For example, a participant generated mediator for the pair *wingu-cloud*, might be *wing*. Participants were required to recall the mediating information at each restudy opportunity. In the test-restudy practice condition, participants were given three cued recall attempts in addition to the restudy trials, to recall the target from the cue. At final test one week later, results showed participants retrieved more target words and mediators in the test-restudy condition than the restudy condition. This was suggested to be due to memory for mediators being enhanced in the test-restudy condition, in comparison to the restudy condition. Furthermore, the test-restudy condition was thought to make better use of effective mediators between the cue and target. For example, in the test-restudy condition ineffective mediators would be more likely to be upgraded for more effective mediators following a retrieval error.

In a test of this view of elaboration, work by Carpenter has found that semantically related items with a greater number of mediators associated with them benefit more from testing than items with fewer mediators (Carpenter, 2011; Carpenter & Yeung, 2017). This work suggested a central role for mediators in the mechanisms of the testing effect, the larger the network of mediators the greater the benefit of testing. However, recent work that examined a more explicit use of mediators in the learning of semantically associated word pairs, found that being able to recall the mediator at final test did not influence the magnitude of the testing effect (Cho, Neely, Brennan, Vitrano, & Crocco, 2017). This suggests that the role of mediators in the testing effect might not be as straightforward as previously suggested by Carpenter and colleagues, yet further work is required to ascertain the use of mediators in the phenomenon of the

testing effect.

While mediating information like keyword mnemonics have shown promise for boosting memory for unassociated concepts, the benefit of utilising keyword mnemonics with retrieval practice has not been fully explored in relation to the direct effects of testing. As Pyc and Rawson's study included feedback in the form of test-restudy trials and we already know testing is more efficient when accompanied with feedback (Rowland, 2014), further work is required to disassociate the benefit of mediating information to retrieval practice, to the benefit to retrieval practice with accompanying feedback. Due to the fact that this has the potential to provide a welcome learning outcome in linking unfamiliar concepts to existing knowledge, more work is required to assess the utility of using mediating information in the learning of unfamiliar concepts.

Although this section has outlined some of the ideas of how elaboration could play a role in the testing effect, it is still not clear to what extent it does or exactly what form elaboration takes and in what way it can be beneficial. However, as already outlined there is a consistent idea that has been explored in relation to retention, which is that meaningful processing could be a relevant aspect to the testing effect. Relatively few studies have explored this concept in any detail, or in the specific ways that have allowed greater understanding of the direct effects associated with retrieval practice. Below I will reiterate where meaningful processing could be relevant to the testing effect and in what ways it will be further explored herein.

1.5 Focus on Meaningful Processing

The evidence reviewed so far repeatedly suggests that meaningful processing could be important to the testing effect. As already outlined at the start of the chapter, meaningful processing will be thought of as processing with great value or significance and will be explored here in relation to how processing items in ways that change their value or significance impacts the magnitude of the testing effect. This will be assessed from a number of different perspectives already detailed.

Experiments 1-4 will assess meaningful processing by examining the extent to

which differences in the semantic relatedness of studied information benefits from retrieval practice. In experiments 1-3, an initial exploration of meaningful processing will focus on the evidence for the elaborate retrieval hypothesis (ERH) (Carpenter, 2009). As already highlighted, this work has not received much attention based on its original form, yet has been often cited as a possible explanation for the testing effect. Therefore, experiments 1-3 aim to test the strength of the original evidence and extend it in line with the robust test-delay interaction effect seen in the literature (Adesope et al., 2017; Rowland, 2014). In experiment 4, a different iteration of the importance of elaboration to the benefit of testing will be examined, the mediator effectiveness hypothesis, which suggests that retrieval practice is more beneficial for items that exploit meaningful semantic networks (Pyc & Rawson, 2010). Crucially, this work has not explored how direct effects of mediation impact the links formed between unrelated concepts. Experiment 4 will explore this.

Experiments 5 and 6 will examine the impact of meaningful processing on the testing effect based on the structural coherence of the study materials. As already highlighted, prose consistently produces greater testing effects (Adesope et al., 2017; Rowland, 2014), suggesting organisation of study materials is significant for testing effects. Yet work in the testing effect literature has not directly assessed this. However, this concept has been explored in an analogous area of research, retrieval induced forgetting, with promising results. Retrieval induced forgetting assesses whether items not retrieved, but related conceptually to items that are retrieved, during a given retrieval episode are detrimentally impacted at final test. Empirical work in this area has shown that items that are associated (M. C. Anderson & McCulloch, 1999) and organised (Chan, 2009) can benefit from retrieval-induced facilitation, in comparison to items that are not associated and not organised. These items are thought to undergo less retrieval-induced forgetting than items that are less meaningfully associated or organised. Crucially, this work is thought to apply to the testing effect (Chan, 2009), but has not utilised a restudy control task. Therefore, in assessing whether structural properties are significant to testing effects this work is further followed up in experiments 5

and 6.

Experiments 7-10 explore two aspects of meaningful processing, firstly by comparing different retrieval practice tasks with appropriate restudy controls. As already highlighted studies have looked to compare what happens during the retrieval practice task from a largely applied perspective, based on which practice task leads to the largest testing effect. Results indicate that different practice tasks influence the testing effect to different extents (Greving & Richter, 2018; Rowland, 2014), but the mechanisms of these differences remain elusive. Studies have found however that what happens during the retrieval practice task can influence what information is recalled (Brewer et al., 2010) and when retrieval is more relational broader benefits in memory are seen (Johnson-Laird et al., 1978).

This has been assessed more directly in the transfer testing effect literature, which assesses more meaningful learning outcomes. For example, rather than asking participants to recall a particular word that was previously seen and tested, the final test might rephrase the question to ask for a different word seen, or ask that knowledge tested during the initial test be applied to a novel problem. In this way transfer testing effects assess the broader application to learning of the testing effect. Work in this area has shown that differences in the retrieval practice task indicate differences in the magnitude of the transfer testing effect. For example, when the retrieval practice task encourages greater depth of processing through focused retrieval, or when the retrieval task utilises increased elaboration, then improvements in transfer learning have been seen (Butler, 2010; Endres et al., 2017; Hinze et al., 2013). However, these studies once more suffer from utilising inadequate restudy control tasks. Therefore experiments 7-10 address meaningful processing based on differences in retrieval practice tasks when the restudy task is matched and the final test requires both direct retention and transfer learning.

It is important to also note, that the key findings outlined earlier in this chapter will inform the perspective that this investigation takes. As already highlighted, much of the work on the testing effect has taken an applied approach. From this viewpoint,

questions like how much can retrieval practice boost learning in relation to some other study strategy and what conditions make testing most effective have been in focus. This approach has been necessary to date, however in an effort to focus the current investigation on the mechanistic properties of the testing effect in light of the key findings and in relation to meaningful processing, the experiments enclosed will: 1) Assess the direct effects of testing, this is testing without the accompaniment of multiple feedback opportunities. 2) Assess the impact of testing under design conditions that boost the effect, namely with a delay of more than one day to final test and with a between-subjects manipulation of the test and restudy conditions. By using this approach consistently, an ability to compare the studies based on these similarities will be established and further conclusions about the mechanistic properties of the effect should be gleaned.

Chapter 2

Revisiting the foundations of the Elaborate Retrieval Hypothesis

Chapter one identified one area of meaningful processing that is necessary to explore, which is how differences in the relational properties of items can contribute to the magnitude of the testing effect. The elaborate retrieval hypothesis suggests that items that require more effort to retrieve, based on the links that are able to be formed between them, should benefit more from retrieval practice. The current chapter revisits the formative results from the ERH to address the extent to which differences in meaningful processing based on elaborate retrieval, relates to the testing effect. Three experiments in this chapter will explore this concept, by extending Carpenter's original work in line with more recently established phenomena in the field.

2.1 Introduction

In starting to examine how meaningful processing might impact the testing effect, first, meaningful processing based on the relatedness of items being studied will be examined. Direct manipulation of meaningful properties of the study materials has not received much attention in relation to the testing effect (Karpicke, 2017). However, a particular view of elaboration theory, the elaborate retrieval hypothesis of the testing effect, suggests that meaningful aspects of the study materials might be key to testing effects. The elaborate retrieval hypothesis, as outlined by Carpenter (2009) and highlighted in chapter one, has garnered much attention over the last eleven years. At the time of writing it has been cited 475 times and many of these citations have endorsed the principles of elaborate retrieval as an explanation for the testing effect. With its foundations in theoretical and empirical work that underpins many aspects of memory research, the elaborate retrieval hypothesis, to its merit, holds a great deal of intuitive appeal. Influences in this area stretch back a long way and have evolved over several

decades (Roediger, Gallo, & Geraci, 2002). These studies have shown that processing items during encoding in multiple ways through multiple *levels* increases later memory (Craik & Tulving, 1975; Einstein et al., 1990; Willoughby et al., 1994).

The description that Carpenter initially gives is rooted in levels of processing through semantic associate networks, but also incorporates elements of effortful retrieval in hypothesising the benefits associated with increased elaboration during retrieval. In this way items that benefit from deeper levels of processing, through activation of an existing strong association network for example, might not require the same level of elaboration during retrieval practice and therefore not benefit in the same way from retrieval practice, as items that are subject to shallower levels of processing. This perspective suggests that both how the materials are encoded, or the level of processing during the study task, and the subsequent retrieval processes associated with the learning of this material is important to the testing effect. This is something that has been under-explored (Karpicke, 2017) and thought to be important for understanding retrieval processes (Tulving & Thomson, 1973; Van Gog & Sweller, 2015).

In 2009, Carpenter proposed the elaborate retrieval hypothesis as an explanation for the testing effect. This perspective examined whether increased memory search as a proxy for more elaboration during the retrieval process could lead to a greater benefit of testing. To explore this, strongly and weakly associated word pairs were used as study materials. Participants completed a typical testing effect paradigm, whereby a list of word pairs were studied once by all participants, the list was then either restudied, or tested, before participants completed a final test on the list. In this instance test practice involved showing participants the cue word learned during study practice and testing whether participants could recall the target word it was paired with. The final test was free recall of the target words. In this experiment, the strong and weak associate word pairs represented a theorised difference in the amount of elaboration required to retrieve each target word. The weaker associates were thought to require a more elaborate memory search, based on the weaker processing during encoding. Subsequently this more elaborate memory search activates more helpful memory links

that are utilised by the learner at a later retrieval point.

The hypothesis was supported across two experiments with four key pieces of evidence. The two experiments differed slightly in design which is worth detailing further here. In experiment one the design was fully within-subjects; participants learned 6 lists of 8 word pairs, three of which were learned via test practice and three via study practice. In each list the number of strongly and weakly associated word pairs were matched and randomly presented. In experiment two the design was fully between-subjects, each participant learned one list of 48 purely strongly or weakly associated word pairs, through either restudy or test practice, giving four conditions.

There are four key findings in support of the elaborate retrieval hypothesis given in this paper. Firstly, results from experiment one showed an interaction between the associative strength of the word pairs and the amount of retrieval at the initial test compared to the final test, showing that the weaker associates demonstrated less forgetting between the initial test and the final test than the stronger associates. This finding was also replicated in experiment two. Secondly, evidence was provided based on the final test data, whereby the association strength of the word pairs (strong, weak) interacted with the type of practice utilised to learn them (restudy, test). This analysis is the testing effect analysis. As such in experiment one, the testing effect analysis also demonstrated that weaker associates benefited more from retrieval practice than stronger associates, when compared to equivalent items that had been restudied. However, the testing effect analysis did not show the same result in experiment two. With an alternative form of analysis however, termed conditional analysis, evidence was shown for the testing effect interaction in experiment two. The conditional analysis was computed as a proportion of correctly retrieved items during the retrieval practice task. This form of analysis is suggested based on the idea that retrieval practice is maximally beneficial when the item has been successfully retrieved (Runquist, 1983), consistent with the bifurcated distribution account (Kornell et al., 2011). When participants had correctly retrieved the item during the practice phase, then the proportion of weaker associate pairs retrieved at the final test was greater in comparison to their restudy equivalents

than the stronger associate pairs.

Carpenter further demonstrated two pieces of evidence based on ratings and response times collected during the task, that suggested that elaborate retrieval could explain the benefit found. The weaker pairs were rated as less related and took longer to retrieve during the retrieval practice phase. This suggested that the weaker associates required a longer, more elaborate memory search than stronger associates, that were quicker and therefore easier to retrieve. The weaker associates were therefore more likely to benefit from the greater amount of information being activated during this longer memory search during a later retrieval attempt.

This evidence taken together seems to be compelling evidence for a benefit for longer, more elaborate retrieval leading to a greater retrieval benefit. However, these results are yet to be directly followed up. In attempting to follow-up these results in exploring the contribution of meaningful processing in the study materials to the testing effect, potential reasons for the less robust evidence found in experiment two will be further explored. One reason for the comparatively lower retrieval rates shown in experiment two than experiment one, could be due to the longer list format of experiment two combined with the final free recall test, which resulted in lower accuracy rates. As free recall and cued recall final tests show mostly equivalent effect sizes for the testing effect (Rowland, 2014), it would seem sensible to instead of giving a free recall final test, giving a cued recall final test. This should boost final retrieval scores and therefore also boost the variation in scores to maximise the likelihood of detecting any differences.

A second potential reason for the lack of consistent evidence across the two experiments, could have been that a short delay (5 minutes) to the final test that was utilised. In wider testing effect research a test-delay interaction can be seen, in which the testing effect becomes larger with time (Adesope et al., 2017; Roediger & Karpicke, 2006b; Rowland, 2014). The magnitude of the effect is typically around 50% greater when the results extend beyond 1 day (Adesope et al., 2017; Rowland, 2014). Therefore, a further way that we could examine the efficacy of these results would be to

extend the findings from this study, to include a greater delay to final test. If test practice is more effective when the target undergoes a longer memory search, indicative of greater elaboration occurring, then we would expect the testing effect to reflect this greater benefit for the weaker associates as the delay to final test increases.

While the evidence from Carpenter's two experiments show some inconsistencies, addressing the reasons for this with amended design features in this chapter will allow for a clearer examination of the contribution of meaningful processing of the study materials to the testing effect. The three experiments given in this chapter will; firstly replicate Carpenter's findings with similar materials and design (experiment 1), secondly extend this initially with a final cued recall test (experiment 2) and finally, to maximise the magnitude of the testing effect with these materials and therefore maximise the likelihood of finding supporting evidence for the elaborate retrieval hypothesis, both a final cued recall task and an increase in delay to the final test will be utilised (experiment 3).

2.2 Experiment 1

Experiment one is designed to replicate the findings from Carpenter (2009), experiment 2. Experiment 2 was chosen for two reasons. Firstly, the utility of the testing effect in higher education and self-testing is likely to be greater if tests can be employed following the study of a good amount of learning materials. Studying with particularly short lists (Carpenter, 2009, experiment 1) represents a somewhat artificial examination of the effect. Secondly, there are some known artifacts associated with learning from shorter mixed lists. For example, when lists to be learned are made of a mixture of high and low frequency items (with 16 items or fewer), a free recall benefit is seen for low frequency items (McDaniel & Bugg, 2008; Ozubko & Joordens, 2007). This is thought to be because low frequency items enjoy greater encoding, perhaps due to their noticeable distinctiveness relative to the high frequency items in short lists. Therefore replicating with longer lists was felt to be a more robust and more generalisable design to use. A similar design to that reported in Carpenter was utilised, adopting a list length of 40 associated word pairs and a within-subjects design. With the changes

made to the design, experiment 1 ensured that results would be comparable to Carpenter's before attempting to extend the results to include further amended design elements in experiments 2 and 3. In addition to a replication, the initial experiment also explored whether the inclusion of feedback as part of the retrieval practice task would influence results in support of ERH. This was due to the fact that elaboration effects have been found to be boosted elsewhere when designs included feedback (Pyc & Rawson, 2010), therefore to maximise the likelihood of replicating the interaction, experiment 1 included a manipulation of feedback.

2.2.1 Methods

Participants and Design

Participants were students at the University of Plymouth (N = 60), aged between 18 and 30 years (M = 20.02 years, SD = 1.84), 83% female. Participants took part in the study for course credit or were paid for their time at £8 per hour, £2 for each 15 minutes.

Experiment 1 utilised a 2 (practice task) x 2 (association strength) x 2 (test type) mixed design, with practice task (restudy, test) and association strength of the word pairs (strongly associated, weakly associated) as within-subjects factors and test type (test only, test with feedback) as a between-subjects factor. Practice task was counterbalanced by subject, participants allocated to the restudy practice task in the first list, completed the second list as test practice and vice versa. Therefore, each participant completed both a study list and a test list and each participant was randomly allocated to test type as either test only or test with feedback.

Sample Size Calculation

To determine the sample size, the calculation was based on finding a medium interaction term, where Carpenter had found large interactions based on the time phase analysis in both experiments. Calculating in G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) based on a medium effect size, η_p^2 = 0.06, with increased power to detect this effect at .95, gave a total sample of 54 participants, based on a 2 x 2 within-subjects

design.

Materials

The word pairs used were very similar to those used in Carpenter's study, with some minor changes. Firstly, two targets were matched to the same cue as opposed to matching two cues to the same target. This ensured that the association network of the cues was matched for both strongly and weakly associated pairs. It also ensured that the strongest mediator of the weakly associated pairs was controlled for. The word pairs consisted of a noun cue word that was taken from the MRC Psycholinguistics database (Wilson, 1988), was between 5 and 7 letters long and had a frequency value of between 20 and 100 per million. All cues were the same for weak and strong associates, meaning that imageability, concreteness and familiarity were the same for the cue words across the two sets. Both weak and strong target words that were linked to each cue were constrained based on their association strength to the cue word. Both the strongly and weakly associated target words in each pair were taken from Nelson's association norms (Nelson, McEvoy, & Schreiber, 1998). The associate strength of each target was calculated based on survey data, for how often participants mentioned each of the forward associates as a percentage. Each noun target was between 5 and 8 words long and each strongly associated target was the strongest associate of the cue word and ranged in value from .15 to .54 with an average associative strength of .32. Each weakly associated target was either the weakest associate, or one of the weakest associates if several words were of equal association strength. Weakly associated targets ranged in value from .01 to .09 with an average strength of .014 and were not a forward associate of any of the remaining cue words. Each participant saw a total of 82 word pairs, consisting of 2 practice items followed by 2 lists of 40 word pairs. For each participant each list consisted of 20 randomly generated weakly associated pairs and 20 randomly generated strongly associated pairs. Each cue was only viewed once across both lists for each participant. All participants saw the same two practice items, which were both strongly associated pairs. A new random presentation

of each list of 40 word pairs was given for the study and practice phases. An example of a strongly associated cue-target pair is *barrier-wall*, with an example of the same cue, with a weakly associated target pair is *barrier-bridge*. The full list of cue-target pairs can be found in appendix A.1.

Procedure

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. Participants were tested either individually or in groups of up to six people. For the duration of the experimental session, participants sat at a partitioned desk with their own PC. Participants wore headphones throughout the task. The presentation of the two lists to be learned followed the same procedure, each with a study phase, a practice phase and a final test phase. The study and practice phases were presented in E-Prime 3.0 software (Psychology Software Tools, Inc, Pittsburgh, PA, 2016) and the final test phase was completed with pen and paper. Participants received two practice items prior to starting the first list, which included an example of both a study and a test trial specific to the condition the participant had been allocated to, either test only or test with feedback. The practice items consisted of an example of two strongly associated word pairs, the cue-target associations were the same across participants. Participants were informed that the types of trials for the practice phase could be either restudy practice trials or test practice trials, so participants did not know in advance of the practice phase which trials they would have on each list. This was done in order to minimise attention bias to either of the lists.

For the study phase all forty word pairs in list one were presented; each word pair was presented for 4 seconds. After 4 seconds the question, *To what extent are these items related?* appeared on the screen. This question was presented to ascertain whether participants were able to detect the differences in association strength between the strong and weak associate pairs. Participants were instructed to make a response on the keyboard between 1 and 5; 1 = unrelated and 5 = very related. After participants made their response a 500ms blank intertrial interval preceded presenta-

tion of the next word pair.

Following presentation of all 40 items on the first list in the study phase, participants complete a two minute task of either a number search or a suduko puzzle. A numerical filler task was chosen for the break, to minimise the likelihood of participants rehearsing something associated to the studied materials. This was presented on a double sided piece of paper and participants were instructed to complete this as they wished whenever a break was indicated. A tone was played through the headphones once the two minutes had elapsed.

The practice phase followed the study phase for each list, participants saw the same 40 word pairs again in a new random order, either as restudy practice items, with the same presentation format as in the study phase (see above), or as test practice items. Test practice items were either as test only or test with feedback. In the test only condition, participants saw the cue word, with the prompt. *Can you remember the target word?* Participants were instructed on screen to either press the space bar and type in the word that they recalled or to press the space bar and type in *no* if they could not remember the target word. Once participants had entered a word they pressed enter for the next practice trial. For participants in the test with feedback condition, once they had pressed enter they saw the word pair on screen again for 3 seconds. Each trial was followed by a 500ms blank intertrial interval before the next cue word was presented.

Following presentation of the practice phase, participants completed a five minute task, during this time participants again completed either a number search or a suduko puzzle.

All participants then completed the final test phase for list one. The final test was a free recall test for the words pairs that had been studied for list one and was completed with pen and paper. Participants were instructed to recall as many word pairs as they could, they were also instructed to write down any individual words that came to mind for which they could not remember the correspondingly paired word. Participants were instructed to recall all word pairs and words in order not to bias their attention on the

second list. However, only successful target recall in a retrieved word pair was counted as successful retrieval. Participants were given six minutes to complete the free recall task for each list.

Following completion of the final test phase on list one, participants completed a 5 minute task, of a number search or suduko puzzle, before proceeding to list two. List two followed the same procedure as list one, except with the alternative practice task (if the first list was restudy practice, the second list was test practice and vice versa). After the final test phase was completed for list two, participants were debriefed and thanked for their time. In experiment 1, the procedure took approximately 45 minutes and was completed in one session.

2.2.2 Results

All analyses were computed in JASP (JASP Team, 2020) and replicated the main analyses reported in Carpenter (2009), detailed in the introduction to this chapter. All frequentist analyses where appropriate are given with the results of bayesian equivalent analyses. Descriptive statistics are given for the main effects of interest in table 2.1.

Coding Responses

Items were coded blind to condition. Plurals incorrectly present or absent, obvious spelling mistakes and two letter changes to make up correct words (but not another word) were coded as correct. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible.

Ratings and Response Times

First analysed, was whether there were differences in the judgements of the relatedness for the strongly and weakly associated word pairs. A paired samples t-test revealed that in line with Carpenter's results, the strongly associated word pairs (M = 4.15, SD = 0.37) were judged to be more related than the weakly associated word pairs (M = 3.49, SD = 0.44), $t(59) = 21.12$, $p < .001$, $d = 2.73$, $BF_{10} > 150$ (8.536e+25). In addition, t-tests were computed comparing response times during the initial test phase

Table 2.1
Initial and Final Test Accuracy in Experiments 1-3 as a Function of Practice Task and Association Strength of Word Pair

	Initial test		Avg	Final test		Avg
Practice task	Strong	Weak	IT	Strong	Weak	FT
			Experiment 1 (*n*=60*)			
Restudy	n/a	n/a	n/a	.23(0.10)	.23(0.12)	.23(0.09)
Test only	.80(0.40)	.63(0.25)	.71(0.18)	.29(0.05)	.25(0.05)	.27(0.09)
Test with FB	.83(0.45)	.66(0.20)	.75(0.14)	.29(0.10)	.30(0.05)	.29(0.07)
Test Avg	.82(0.14)	.64(0.20)	.73(0.16)	.29(0.11)	.27(0.11)	.28(0.08)
			Experiment 2 (*n*=30)			
Restudy	n/a	n/a	n/a	.94(0.10)	.85(0.14)	.89(0.11)
Test	.89(0.16)	.80(0.12)	.84(0.12)	.90(0.11)	.78(0.17)	.84(0.12)
			Experiment 3 (*n*=30)			
Restudy	n/a	n/a	n/a	.47(0.21)	.23(0.16)	.35(0.18)
Test	.82(0.17)	.69(0.20)	.76(0.17)	.60(0.22)	.38(0.20)	.49(0.20)

Note. The values represent mean percentages of target words recalled, SDs given in parentheses. *Test condition was between-subjects, for each test condition *n*=30.

between strongly and weakly associated pairs. In line with Carpenter's results, both correct responses (strong, M = 1803, SE = 54.88; weak, M = 2132, SE = 78.38) and all responses (strong, M = 2042, SE = 69.14; weak, M = 2740, SE = 129) showed quicker response times for the strongly associated word pairs compared to the weakly associated word pairs (correct responses, $t(59) = -4.89$, $p < .001$, $d = -0.63$, $BF_{10} > 150$ (2187); all responses, $t(59) = -6.07$, $p < .001$, $d = -0.78$, $BF_{10} > 150$ (141079)). These findings suggest that participants take longer to search memory for the weakly associated word pairs during retrieval practice, which is consistent with the claims of the elaborate retrieval hypothesis.

Main Analyses

Next, the three main findings relevant to this replication are assessed. The first being the interaction between weakly and strongly associated word pairs, from initial practice test accuracy to final target word recall, this will be termed test phase comparison. The second finding is an interaction for the testing effect, between practice task and association strength, which detects whether a testing effect interaction is present. The third is an interaction for the testing effect where accuracy is recorded based on whether items were accurately retrieved during the initial test. This will be termed conditional analyses. In each case a greater benefit for the weak associate pairs is expected than for the strong associate pairs.

Test phase comparison. Carpenter (2009, Experiment 2) found that for initial test performance, strongly associated word pairs showed better recall than weakly associated word pairs. Furthermore, this pattern interacted with the results at final test, such that there was no longer an advantage for the strongly associated word pairs on the final test. To test whether the current results are consistent with these findings, a 2 (test phase; initial test, final test) x 2 (association strength; strong, weak) x 2 (feedback; present, absent) mixed ANOVA was conducted, with test phase and association strength as within-subjects factors and feedback as a between-subjects factor.

The same pattern found by Carpenter is reported here and is depicted in figure 2.1. There was a strong effect of test phase, with initial retrieval accuracy (M = .73, SD = 0.16), being greater than final test accuracy (M = .31, SD = 0.09), $F(1,58)$ = 746.31, $p < .001$, $\eta_p^2 = .93, BF_{10} > 150$ (1.705e+65). This was to be expected with the difference in test format between the initial (cued recall) and final test (free recall) phases.

There was evidence for a main effect of association strength, with strong associates (M = .56, SD = 0.10) retrieved more often than weak associates (M = .30, SD = 0.12), $F(1,58)$ = 49, $p < .001$, $\eta_p^2 = .46, BF_{10}$ = 5.49. This analysis was not previously reported by Carpenter. A significant interaction between test phase and association strength was found, $F(1,58)$ = 47.75, $p < .001$, $\eta_p^2 = .45, BF_{10} > 150$ (273586.44) (in-

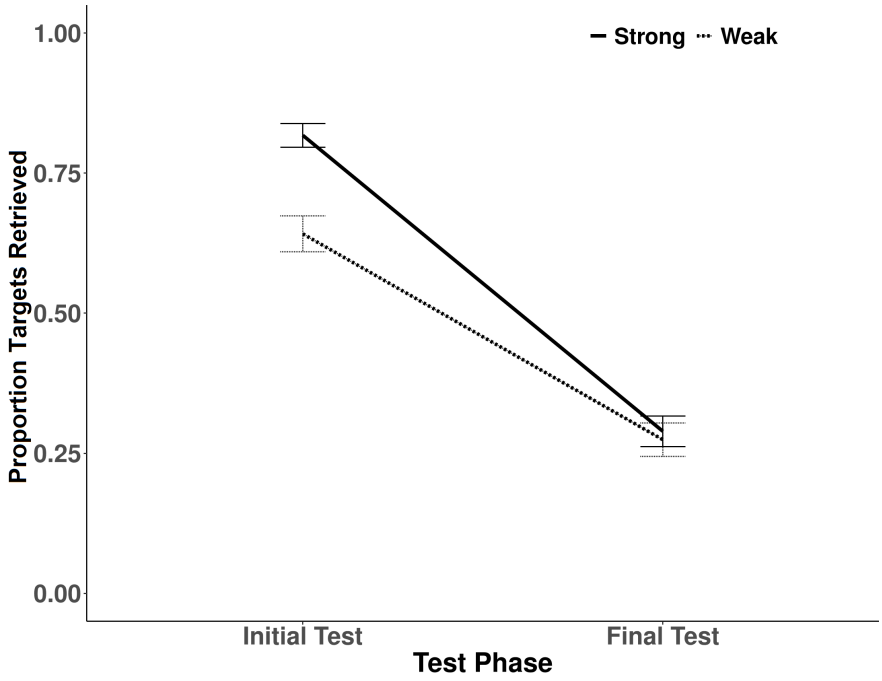

Figure 2.1. Mean target retrieval as a function of test phase and association strength in experiment 1. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

teraction bayes factor terms in this text are given for matched models effects in JASP as suggested by van den Bergh et al., 2020). There was no main effect of feedback and this did not interact with any effects of interest, Fs < 1.40. Follow-up one-way ANOVAs revealed this to be in the predicted direction, with differences found between strong associate and weak associate target retrieval accuracy at initial test, $F(1,58) = 104.72$, $p < .001$, $BF_{10} > 150$ (4.973e+11), but not at final test, $F(1,58) = 0.65$, $p = 0.42$, $BF_{10} = 0.26$. A directional bayesian t-test assessed the evidence in favour of the elaborate retrieval hypothesis, that the rate of decay for the weak associate pairs was less than the rate of decay for the strong associate pairs, set with a default prior. Difference scores were calculated between initial test accuracy and final test accuracy (initial test score - final test score) for both the weak (weak difference score) and strong associates (strong difference score). The bayesian t-test assessed the strength of evidence that the *weak difference score* $<$ *strong difference score*, results revealed extreme evidence in favour of this prediction, $BF_{10} > 150$ (5.601e+6).

Testing effect. To examine whether final test performance would reflect this retrieval practice advantage for weaker associates, a further 2 (practice task; restudy, test) x 2 (association strength; strong, weak) x 2 (feedback; present, absent) mixed ANOVA was conducted, with practice task and association strength as within-subjects factors and feedback was a between-subjects factor. Consistent with Carpenter's results, a main effect of practice task was found, with test practice (M = .28, SD = 0.08) resulting in better retrieval of targets than restudy practice (M = .23, SD = 0.09), F(1, 58) = 17.43, $p < .001$, $\eta_p^2 = .23, BF_{10} > 50$ (160.57). There was no effect of association strength, with equivalent retrieval of targets seen for the strong associates (M = .26, SD = 0.09) and the weak associates (M = .25, SD = 0.09), F(1,58) = 0.37, p = .55, $\eta_p^2 = .006, BF_{10}$ = 0.16. Furthermore, there was no interaction between practice task and association strength, F(1,58) = 0.33, p = .57, $\eta_p^2 = .006, BF_{10}$ = 0.23. These results are depicted in figure 2.2. Instead evidence for the lack of an interaction is given based on the bayes factor reported, whereby the evidence suggests that H_0 is 4.35 times more likely than H_1. This suggests positive or substantial evidence in favour of the null hypothesis of no interaction (Jarosz & Wiley, 2014). There was no main effect or interaction with feedback, therefore this factor is not further reported here, Fs < 3.10.

Conditional analyses. Finally, a conditional analysis was conducted, with final test accuracy measured as a proportion of initial test accuracy for the test practice trials, consistent with the analysis completed by Carpenter. A final 2 (practice task; restudy, test) x 2 (association strength; strong, weak) x 2 (feedback; present, absent) mixed ANOVA was computed, with practice task and association strength as within-subjects factors and feedback was a between-subjects factor. Results mirrored the main results found in the testing effect analysis, with a main effect of practice task, whereby target retrieval for test practice trials (M = .31, SD = 0.09) was greater than target retrieval for restudy practice trials (M = .23, SD = 0.09), F(1,58) = 33.49, $p <$ .001, $\eta_p^2 = .37, BF_{10} > 150$ (136340.86). Again, no main effect of association strength was found, with strong (M = .27, SD = 0.09) and weak associate target retrieval (M =

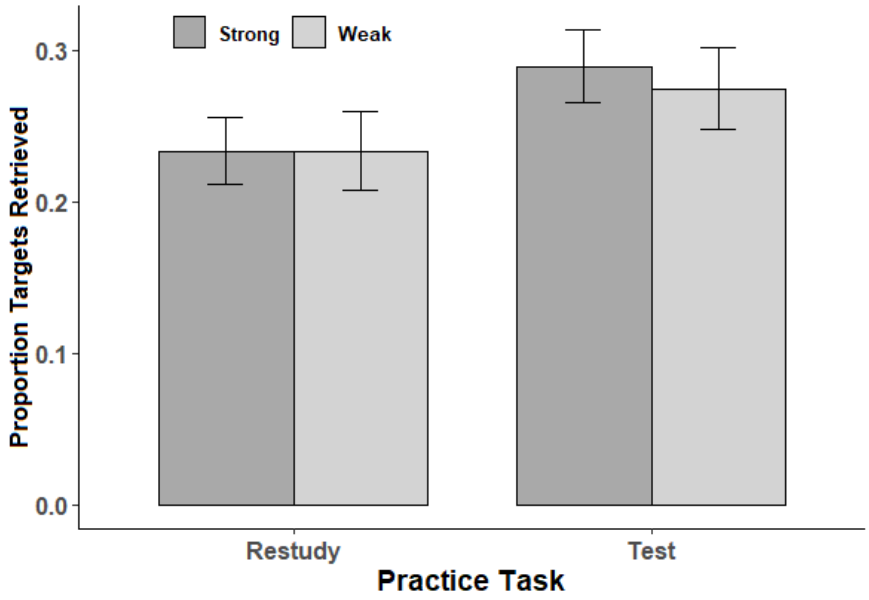

Figure 2.2. Mean target retrieval at final test as a function of practice task and association strength in experiment 1. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

.28, *SD* = 0.09) showing equivalent recall rates, $F(1,58) = 0.57$, $p = .45$, $\eta_p^2 = .01, BF_{10} = 0.17$. Once more there was no evidence for an interaction, $F(1,58) = 0.42$, $p = .52$, $\eta_p^2 = .007, BF_{10} = 0.24$. Instead, there is again positive or substantial evidence in favour of the null hypothesis, the evidence suggests that H_0 is 4.17 times more likely than H_1. Results for feedback showed an interaction with practice task, but results were not in line with the hypotheses. [1]

2.2.3 Discussion

The results from experiment 1 replicated the main findings from Carpenter (2009, experiment 2), suggesting that the materials used here are a good match to those previ-

[1] There was a main effect of feedback found on the conditional analysis, with feedback absent resulting in greater accuracy (*M* = .30, *SD* = 0.06) than when feedback was present (*M* = .25, *SD* = 0.08), $F(1,58) = 7.76$, $p < .01$, $\eta_p^2 = .12, BF_{10} = 3.07$. Feedback further interacted here with the practice task, $F(1,58) = 4.76$, $p = .03$, $\eta_p^2 = .08, BF_{10} = 1.46$. Follow-up one-way ANOVAs here based on restudy compared to conditional test when feedback was present versus when it was absent, showed that this was due rather unexpectedly to there being a larger difference between restudy and conditional test scores when feedback was absent (Restudy: *M* = .24, *SD* = 0.08, Test: *M*= .35, *SD* = 0.08), $F(1,29) = 30.68$, $p < .001$, $\eta_p^2 = .51, BF_{10} > 1000$) than when it was present (Restudy: *M* = .22, *SD* = 0.10, Test: *M* = .27, *SD* =0.08), $F(1,58) = 8.82$, $p < .01$, $\eta_p^2 = .23, BF_{10} = 8.23$). Feedback did not further interact with any of the factors of interest.

ously used by Carpenter and suitable to utilise in the follow-up experiments. The measures of relatedness indicated that participants viewed the stronger associate pairs as more related than the weaker associate pairs. In addition, participants took longer to retrieve the target words from memory for the weaker associate pairs during the initial test practice task. Consistent with the ERH this could indicate that a greater search of memory is being conducted during retrieval practice of the weaker associate pairs.

As with Carpenter's findings, test phase was found to interact with association strength, showing that the weaker associate pairs benefited more from the final free recall test following the cued recall initial test than the stronger associate pairs. However, somewhat surprisingly neither for the absolute accuracy or conditional accuracy analyses did this translate to a larger testing effect at the final test for the weaker pairs. Generally, items that are more difficult to retrieve result in larger testing benefits, therefore it is somewhat surprising not to see that pattern reflected here in the classic testing effect analyses. As already discussed this could be due to the briefness of the delay between practice test and final test potentially masking the differences. In addition, as the final test was free recall and resulted in low final retrieval rates, it might be that offering more cues during the final retrieval test would boost performance and potentially reveal greater differences. Due to the fact that including a delay with a free recall task would likely result in floor effects based on the performance here, the next step is to include a cued recall final test with the same delay to final test of 5 minutes given in experiment 1.

One element of difference between the current and previous findings in the testing effect literature more broadly involves the impact of feedback. Feedback, either through correct answer reveal or additional study time after a response is given, as in the current study, typically enhances the benefits of retrieval practice (Rowland, 2014), although the benefit is not always reported to be large (Adesope et al., 2017; Mulligan et al., 2016). However, in the current study no benefit of feedback was found in relation to the testing effect. It is possible that when initial accuracy is high, feedback is less impactful (Butler & Roediger, 2007). Elsewhere the inclusion of feedback is sug-

gested to prevent the exploration of the direct effects of retrieval (Karpicke et al., 2014), which are those disassociated from any secondary effects linked to the provision of feedback. For example, Kornell, Hays, and Bjork (2009) found that retrieval enhances the subsequent encoding of information, which means both feedback and restudy opportunities following retrieval are not equal to restudy opportunities that do not follow retrieval. Therefore, as testing effects are robustly found in the absence of feedback and feedback has not demonstrated any notable impact on the results of the current study, feedback will be dropped from the remaining experiments in this chapter.

2.3 Experiment 2

Experiment 2 looked to address whether the results of experiment 1 would be replicated and extended with the inclusion of a cued recall, as opposed to free recall, final test.

2.3.1 Methods

Participants and Design

Participants were 30 students at the University of Plymouth, aged between 18 and 35 years (M = 21.0, SD = 4.29), 76.9% female. Participants took part in the study for course credit or were paid for their time at £8 per hour, £2 each 15 minutes (12 participants were paid).

In experiment two a 2 x 2 within-subjects design was utilised, with factors practice task (restudy, test) and association strength (strong, weak). As in the previous experiment, task order was counterbalanced by participant, with an equal number of participants completing the restudy practice task first as those completing the test practice task first. All other list details were the same as in experiment one.

Sample Size Calculation

The sample size for experiment 2 was adjusted based on the evidence from experiment 1. A more conservative estimate of the testing effect found in experiment 1 was made.

It was thought reasonable to expect a large effect η_p^2 = 0.10, based on the previous experiment where a testing effect of η_p^2 = 0.23 was found. Based on this effect size, a G*Power (Faul et al., 2009) analysis suggested 20 participants would be required, to detect this size effect in a 2 x 2 within-subjects design, with power of 0.80. Erring on the side of increased power, this figure was rounded up to 30 participants for experiment 2.

Materials

The word pairs used in experiment 2 were the same as those used in experiment 1, the full details of the study materials can be found in appendix A.1.

Procedure

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. Participants were tested either individually or in groups of up to six people. For the duration of the experimental session, participants sat at a partitioned desk with their own PC. Participants wore headphones throughout the task. In experiment two, participants were required to attend one session which took approximately 45 minutes to complete. All elements of the task were presented using E-Prime 3.0 software (Psychology Software Tools, Inc, Pittsburgh, PA, 2016). Participants learned both lists in succession in experiment 2 before moving on to complete the final test phase (cued recall) after a 5 minute filler task. Due to the cued recall paradigm there was unlikely to be interference related issues that would require immediate testing following the practice phase as was the case in experiment 1. Therefore, both lists were learned first and the final test phase was administered after learning both lists. This format was chosen in order to prevent participants from paying more attention to one item of the pair during the learning of the second list, or adopting a different learning strategy for each list. The final cued recall test phase was completed in the same list order as the practice task and lists were separated by a one minute filler task, which was a number search or sudoko puzzle. Following completion of the session, participants were debriefed and thanked for their time.

2.3.2 Results

As with experiment one all analyses were computed in JASP (JASP Team, 2020). The same analyses were conducted for experiment 2 as in experiment 1. All frequentist analyses where appropriate are given with the results of bayesian equivalent analyses. Descriptive statistics are given for the main effects of interest in table 2.1.

Coding Responses

Items were coded blind to condition. Plurals incorrectly present or absent, obvious spelling mistakes and two letter changes to make up correct words (but not another word) were coded as correct. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible.

Ratings and Response Times

Firstly, as in the previous experiment, relatedness ratings during the initial study phase across both lists for strong and weak associate pairs were compared. A paired samples t-test revealed that the stronger associates were rated as more related (M = 3.99, SD = 0.47) than weaker associates (M = 3.20, SD = 0.48), $t(1, 29) = 15.53$, $p < .001$, d = 2.83, $BF_{10} > 1000$ (4.306e+12).

Secondly, response times during test practice were compared for retrieving strong and weak associate targets. A paired samples t-test revealed that on correct responses the stronger associate targets (M = 1851, SD = 596) were responded to more quickly than weaker associate targets (M = 2222, SD = 701), $t(1, 29) = -4.18$, $p < .001$, d = -0.76, $BF_{10} > 100$ (115.65). Across all responses during the practice task this trend remained consistent, as strong associate targets (M = 2097, SD = 768) were responded to more quickly than weak associate targets (M = 2860, SD = 1211), $t(1, 29) = -3.90$, $p < .001$, d = -0.71, $BF_{10} > 50$ (58.75).

Main Analyses

As with experiment 1 analyses, three main tests relevant to this experiment are examined. The first being the interaction between strongly and weakly associated word

pairs, from initial practice test accuracy to final target word recall. The second finding is the main testing effect interaction for final test target recall between the practice task and association strength. Thirdly, the testing effect interaction is assessed when the test practice accuracy rates are conditional on initial test performance. Across all three findings, a greater benefit for the weak associate pairs compared to the strong associate pairs was expected with the cued recall final test, where the evidence was not overwhelming in experiment 1.

Test phase comparison. The test phase comparison was computed with a 2 (test phase; initial test, final test) x 2 (association strength; strong, weak) within-subjects ANOVA, based on the accuracy scores for target recall on the initial test and the final test. Results revealed no main effect of test phase, with equivalent levels of target accuracy during the initial test (M = .84, SD = 0.12) and the final test (M = .84, SD = 0.12), $F(1, 29) = 2.22$, $p = .15$, $\eta_p^2 = 0.07, BF_{10} = 0.20$. There was a main effect of association strength, with strongly associated targets (M = .90, SD = 0.11) being recalled more accurately than weakly associated targets (M = .79, SD = 0.16), $F(1, 29) = 19.40$, $p < .001$, $\eta_p^2 = 0.40, BF_{10} > 1000$ (1.078e+8). The ANOVA revealed a significant interaction, $F(1, 29) = 8.37$, $p < .01$, $\eta_p^2 = 0.22, BF_{10} = 0.38$, although this was not supported by the evidence given in the bayes factor [2]. This result is depicted in Figure 2.3. Follow-up one-way ANOVAs revealed stronger evidence for differences at final test, $F(1, 29) = 22.15$, $p < .001$, $BF_{10} = 523.8$, between strong and weak associates than at initial test, $F(1, 29) = 15.35$, $p < .001$, $BF_{10} = 57.17$. This result is not consistent with the results from experiment 1, as the differences have increased from initial test to final test, rather than being eliminated by final test. A directional bayesian t-test assessed the evidence in favour of the elaborate retrieval hypothesis, that the rate of decay for the weak associate pairs was less than the rate of decay

[2]Upon further inspection, when the within-subjects frequentist ANOVA was run as between-subjects the interaction effect disappeared, more closely mapping on to the bayesian results. This is thought to be due to the large main effect masking the within-subject variation in the bayesian analysis. The within-subject adjustment that has been ubiquitously adopted for within-subjects frequentist ANOVAs does not yet have a bayesian equivalent. However, the development and adoption of such an adjustment for available software is likely imminent (Nathoo, Kilshaw, & Masson, 2018).

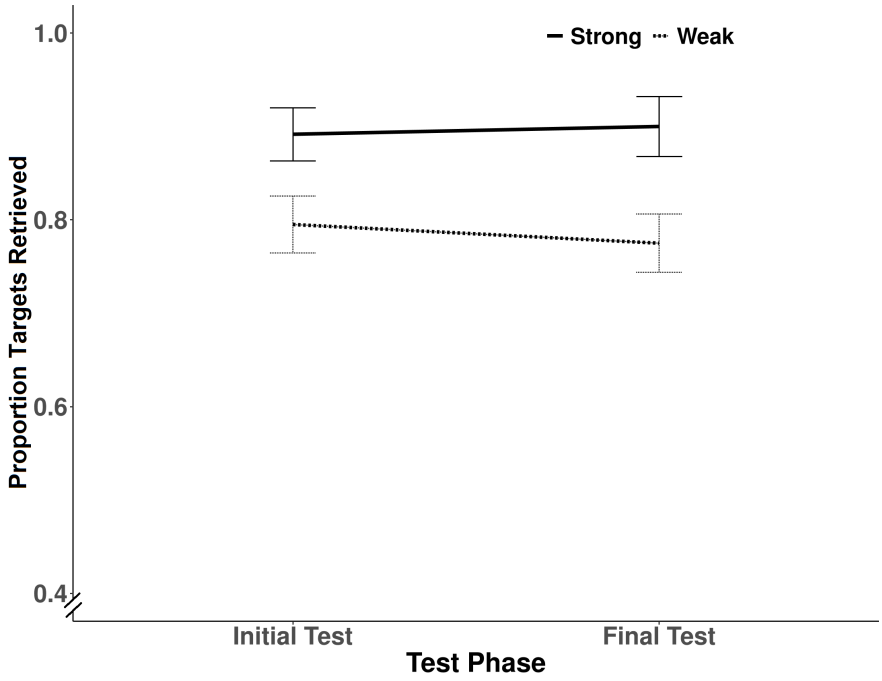

Figure 2.3. Mean target retrieval as a function of test phase and association strength in experiment 2. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

for the strong associate pairs, set with a default prior. As in experiment 1, difference scores were calculated between initial test accuracy and final test accuracy (initial test score - final test score) for both the weak (weak difference score) and strong associates (strong difference score). The bayesian t-test assessed the strength of evidence that the *weak difference score* $<$ *strong difference score*. Results revealed strong evidence in favour of the null hypothesis that strong differences are equal to or smaller than weak differences, BF_{10} = 0.06.

Testing effect. For the main analysis assessing the testing effect, a 2 (practice task; restudy, test) x 2 (association strength; strong, weak) within-subjects ANOVA was conducted on final test target retrieval. There was a main effect of practice task, with restudy practice target accuracy (*M* = .89, *SD* = 0.11) greater than test practice target accuracy (*M* = .84, *SD* = 0.12), $F(1,29) = 9.91$, $p < .01$, $\eta_p^2 = 0.26, BF_{10} = 4.79$. This result indicates no testing effect was found, rather a significant negative testing effect was found, which is a benefit for restudy practice. Results revealed a main effect of

association strength, whereby strongly associated targets (*M* = .92, *SD* = 0.08) were recalled more often than weakly associated targets (*M* = .81, *SD* = 0.14), $F(1,29)$ = 42.5, $p < .001$, $\eta_p^2 = 0.59, BF_{10} > 1000$ (1.058e+6). There was no evidence for an interaction between practice task and association strength, $F(1,29)$ = 1.19, p = .28, $\eta_p^2 = 0.04, BF_{10}$ = 0.40, this bayesian evidence suggests that the H_0 is 2.5 times more likely than H_1, although is not conclusively in favour of the null hypothesis. Results are depicted in figure 2.4.

Conditional analyses. In line with Carpenter's reporting, a conditional analysis of the testing effect result was conducted as per experiment 1. It is worth noting here, that the main results do not support a testing effect as the restudy practice accuracy was greater than the test practice accuracy. However, it is reasonable to assume that if weaker associate networks benefit to a greater extent from retrieval, then this should be reflected in the size of the negative testing effect. In this way, perhaps a reduced deficit for the weaker associates learned through test practice as opposed to restudy practice would be seen than with the strong associates.

A final 2 (practice task; restudy, test) x 2 (association strength; strong, weak) within-subjects ANOVA was computed on the final test target accuracy data. Performance for test practice was conditional on successfully retrieving the item during the initial practice phase. The same main effects were found as the testing effect analysis, in addition to no evidence for an interaction between the two factors, $F(1, 29)$ = 0.73, p = .40, $\eta_p^2 = 0.03, BF_{10}$ = 0.35. This bayesian evidence based on the conditional data, suggests a strengthening of the evidence for the null hypothesis from the testing effect analysis above, making H_0 2.86 times more likely than H_1, however, again this evidence is not thought to be strong enough to be conclusive in favour of the null hypothesis.

2.3.3 Discussion

In line with the results from experiment 1, participants' relatedness ratings and response times during retrieval practice again demonstrated evidence for processing dif-

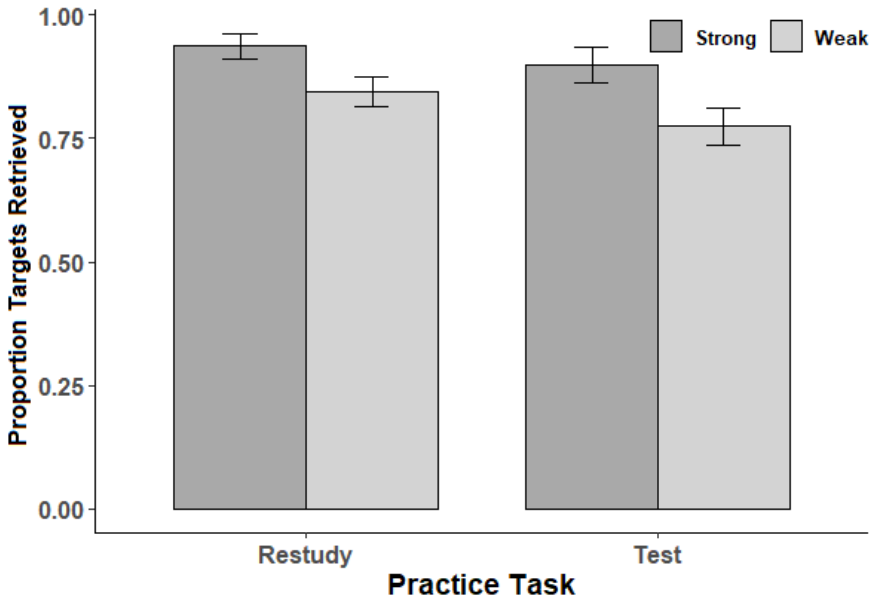

Figure 2.4. Mean target retrieval at final test and association strength as a function of practice task in experiment 2. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

ferences between the strong and weak associates. However, this did not result in a benefit for the weaker associate pairs through testing with a final cued recall test during a delay of 5 minutes. Instead, a benefit was found for the stronger associates rather than the weaker associates in the test phase analysis, based on the results of the frequentist ANOVA. Furthermore, the testing effect analysis did not show a benefit compared to restudy for weak associate target recall. These results, in addition to experiment 1, again shows a lack of evidence for the elaborate retrieval hypothesis of the testing effect.

Caution is required when interpreting these results as a positive testing effect was not found with the immediate final cued recall paradigm. Instead, a negative testing effect is reported, whereby restudy practice accuracy at final test outperformed test practice accuracy. However, this pattern of results, of finding a restudy advantage over retrieval practice, is not uncommon when final test sessions are held immediately (Roediger & Karpicke, 2006b; Rowland, 2014). While previous results might have predicted the lack of positive testing effect in experiment 2 (Roediger & Karpicke, 2006b), positive testing effects are reliably found when an immediate test is utilised although

with smaller effects (Rowland, 2014). Furthermore, previous immediate restudy advantage results have been found in the context of repeat practice blocks (Roediger & Karpicke, 2006b), which the current study did not utilise. In addition, positive testing effects are more likely to be found when the initial test is cued recall (Rowland, 2014), which the current experiment did utilise. Therefore, on reflection there was no concrete evidence that absence of a testing effect, or a negative testing effect would be found here. More broadly, negative testing effects have been reported under certain design conditions when the final test is free recall (Mulligan & Peterson, 2015; Peterson & Mulligan, 2013). However, the predicted mechanisms of this phenomenon are not thought to extend to a cued recall test, which should benefit retrieval practice to a greater extent due to increased cue-target associations and item level processing over restudy. Currently accounts of the testing effect have given little attention to explaining or acknowledging negative testing effects. These results are typically explained based on a lack of delay to final test not revealing the retrieval benefit to forgetting.

The immediate cued recall design of experiment 2 was also practical from an operational viewpoint, in attempting to initially extend Carpenter's findings with minimal change to the original design. It is possible that because there was no advantage for retrieval practice at final test here, no weak associate benefit was revealed despite evidence from the processing metrics of increased elaboration for the weak associate pairs. Therefore, in order to validate the lack of evidence for the elaborate retrieval hypothesis in the results of experiments 1 and 2, it will be necessary to follow-up this experiment with a design that will enable a substantial positive testing effect. Experiment three aims to address this issue, by exploring whether Carpenter's results can extend to the testing effect phenomenon of the test-delay interaction in experiment 3. To explore this experiment 3 will include both a cued recall final test and a delay period of 3-5 days to the final test.

2.4 Experiment 3

Experiment 3 sought to extend the findings of experiments 1 and 2 by utilising a greater delay to the final cued recall test. In particular, we sought to examine whether the advantage for weaker associates could be demonstrated more robustly over a time delay of 3-5 days between the initial test and the final test, as such a delay is observed to boost the effects of testing (Roediger & Karpicke, 2006b; Rowland, 2014). Experiment three took place over two sessions, with the study and practice phases taking place in session one and the final test phase taking place in session two.

2.4.1 Methods

Participants and Design

Participants were 30 students at the University of Plymouth, aged between 18 and 29 years (M = 21.30, SD = 2.67), 76.7% female. Participants took part in the study for course credit or were paid for their time at £8 per hour, £2 each 15 minutes.

In experiment 3, a 2 x 2 within-subjects design was utilised, with factors practice task (restudy, test) and association strength (weak, strong). As in the previous experiments, task order was counterbalanced between participants, with an equal number of participants completing the restudy practice task first as those completing the test practice task first. All other list details were the same as in experiments 1 and 2.

Sample Size Calculation

The sample size for experiment 3 was adjusted based on the evidence from experiment 1, due to the fact that a negative testing effect was found in experiment 2. A more conservative estimate of the testing effect found in experiment 1 was made. Once more it was thought reasonable to expect a large effect η_p^2 = 0.10, based on the previous experiment where a testing effect of η_p^2 = 0.23 was found. Based on this effect size, a G*Power (Faul et al., 2009) analysis suggested 20 participants would be required, to detect this size effect in a 2 x 2 within-subjects design, with power of 0.80. Erring on the side of increased power, this figure was rounded up to 30 participants once more

for experiment 3.

Materials

The word pairs used in experiment 3 were the same as those used in experiments 1 and 2, the full details of the study materials can be found in appendix A.1.

Procedure

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. Participants were tested either individually or in groups of up to six people. For the duration of both experimental sessions, participants sat at a partitioned desk with their own PC. Participants wore headphones throughout session one. In experiment 3, participants were required to attend two sessions, session one took approximately 30 minutes to complete and session two took 15 minutes to complete. All elements of the task were presented using E-Prime 3.0 software (Psychology Software Tools, Inc, Pittsburgh, PA, 2016). Participants learned both lists in the same manner as in experiment 2, however instead of completing a final test phase, after participants had learned both lists in session one, they were instructed to return for a test in session two. The final test was scheduled 3 to 5 days after session one. To ensure adequate recall rates, participants were given a cued recall test in the second session, instead of the free recall test in experiment 1. The cued recall test was completed in the same list order as in the learning session, the lists were separated by a one minute task, in which participants completed a puzzle before continuing on to the final test phase for the second list. Again following completion of the second session, participants were debriefed and thanked for their time.

2.4.2 Results

As with the previous two experiments all analyses were computed in JASP (JASP Team, 2020). The same analyses were conducted for experiment 3 as experiments 1 and 2. All frequentist analyses where appropriate are given with the results of bayesian equivalent analyses. Descriptive statistics are given for the main effects of interest in

table 2.1.

Coding Responses

Items were coded blind to condition. Plurals incorrectly present or absent, obvious spelling mistakes and two letter changes to make up correct words (but not another word) were coded as correct. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible.

Ratings and Response Times

Firstly, a paired samples t-test was conducted on the relatedness ratings for the two types of word pairs. Consistent with the previous two experiments, strongly associated word pairs (M = 4.21, SD = 0.41) were rated as more related than weakly associated word pairs (M = 3.59, SD = 0.57), $t(29)$ = 11.0, $p < .001$, d = 2.01, $BF_{10} > 150$ (1.211e+9).

Response times during the practice test to retrieve correct responses and all responses were compared between the strongly and weakly associated word pairs. Paired samples t-tests revealed only a difference was found across all responses (correct responses, $t(29)$ = -0.95, p = .35, d = 0.17, BF_{10} = 0.29 ; all responses, $t(29)$ = 2.64, p = .01, d = 0.48 , BF_{10} = 3.53), with weaker associates (M = 2424, SE = 150.6) showing greater retrieval times across all word pairs than stronger associates (M = 2068, SE = 116.6).

Main Analyses

Test phase comparison. Firstly, whether the benefit for weak associates from initial test to final test phase was more marked than for strong associates with a delayed cued recall final test was assessed. A 2 (test phase; initial test, final test) x 2 (association strength; strong, weak) within-subjects ANOVA revealed a main effect of test phase, as target recall on the initial test (M = .76, SD = 0.16) was greater than on the final test (M = .49, SD = 0.20), $F(1,29)$ = 100.45, $p < .001$, $\eta_p^2 = 0.78, BF_{10} >$ 150 (2.968e+11). There was also a main effect of association strength, indicating that

the strongly associated word pairs (*M* = .71, *SD* = 0.18) were consistently better recalled at both time points than the weakly associated word pairs (*M* = .53, *SD* = 0.18), $F(1,29) = 52.61$, $p < .001$, $\eta_p^2 = 0.65, BF_{10} > 150$ (4960.19). There was some evidence for an interaction between test phase and association strength ($F(1,29) = 7.21$, $p = .01$, $\eta_p^2 = 0.20, BF_{10} = 0.78$). However, this was wholly not supported by the bayesian evidence. Follow-up one-way ANOVAs revealed greater evidence for differences between strong and weak accuracy at final test, $F(1,29) = 64.94$, $p < .001$, $BF_{10} > 150$ (930626), than initial test, $F(1,29) = 22.83$, $p < .001$, $BF_{10} > 150$ (395.20). A directional bayesian t-test assessed the evidence in favour of the elaborate retrieval hypothesis, that the rate of decay for the weak associate pairs was less than the rate of decay for the strong associate pairs, was conducted with a default prior. Once more, difference scores were calculated between initial test accuracy and final test accuracy (initial test score - final test score) for both the weak (weak difference score) and strong associates (strong difference score). The bayesian t-test assessed the strength of evidence that the *weak difference score* $<$ *strong difference score*. Results revealed strong evidence in favour of the null hypothesis that strong differences are equal to or smaller than weak differences, $BF_{10} = 0.06$. This result is consistent with the results from experiment two and does not support the ERH or original hypotheses. Results are depicted in figure 2.5.

Testing effect. Again, a 2 (practice task; restudy, test) x 2 (association strength; strong, weak) within-subjects ANOVA was computed to examine whether a testing effect was found, based on the target recall accuracy at final test. There was a main effect of practice task, with the testing condition (*M* = .49, *SD* = 0.20) demonstrating greater final test target recall than the restudy condition (*M* = .35, *SD* = 0.18), $F(1,29) = 26.08$, $p < .001$, $\eta_p^2 = 0.47, BF_{10} > 150$ (517.30). There was a main effect of association strength, $F(1,29) = 162.37$, $p < .001$, $\eta_p^2 = 0.85, BF_{10} > 150$ (3.778e+10), with strong associates (*M* = .53, *SD* = 0.20) showing greater final test target recall than weak associates (*M* = .31, *SD* = 0.16). However there was no evidence for an

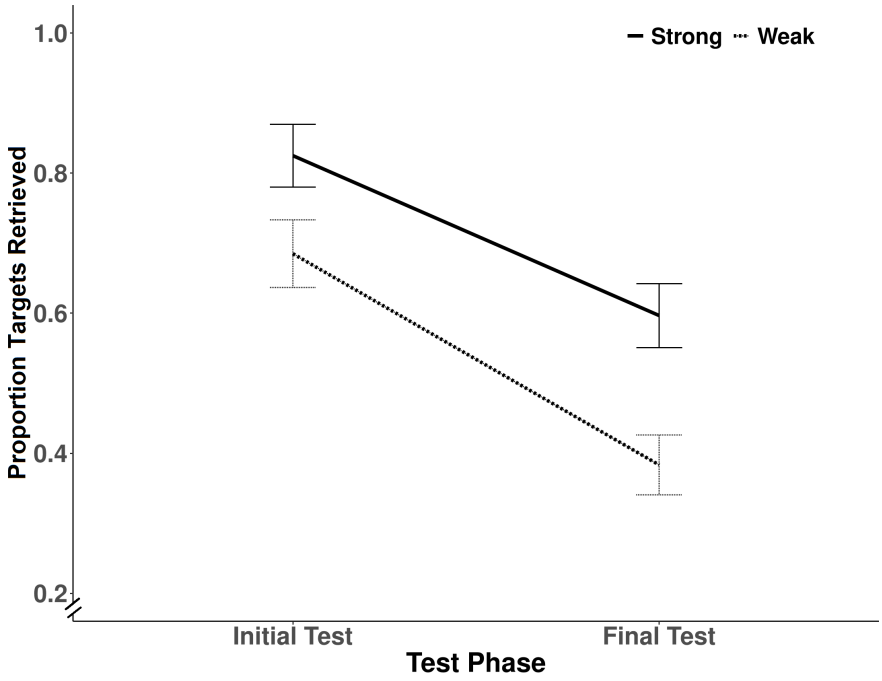

Figure 2.5. Mean target retrieval as a function of test phase and association strength in experiment 3. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

interaction between practice task and association strength, $F(1,29) = 0.40$, $p = .53$, $\eta_p^2 = 0.01, BF_{10} = 0.36$. This bayesian result is approaching conclusive evidence for the null hypothesis, with H_0 being 2.78 times more likely than H_1. This result is depicted in figure 2.6.

Conditional analyses. A final 2 (practice task; restudy, test) x 2 (association strength; strong, weak) within-subjects ANOVA was computed on final test target recall accuracy, whereby final test performance was conditional on successful initial test retrieval. The same main effects were found as in the testing effect analysis, in addition to evidence for no interaction between practice task with conditional test data and association strength on the final test target recall, $F(1,29) = 0.46$, $p = .50$, $\eta_p^2 = 0.02, BF_{10} = 0.30$. With this bayesian result there is conclusive evidence for no interaction between the two factors.

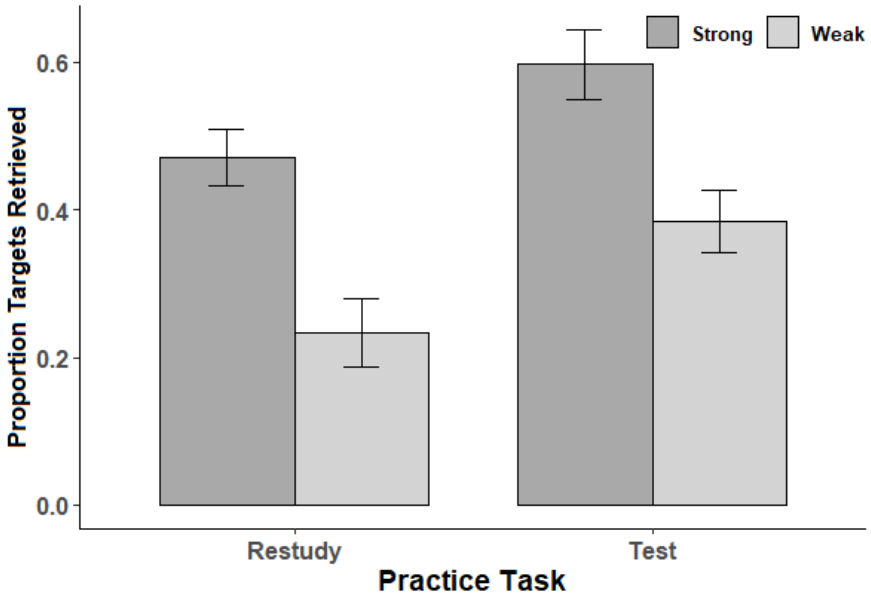

Figure 2.6. Mean target retrieval at final test as a function of practice task and association strength in experiment 3. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

2.4.3 Discussion

In line with the results from experiments 1 and 2, participants' relatedness ratings and response times during retrieval practice further indicated processing differences between the strong and weak associates in experiment 3. However, again there was no evidence that the increased difficultly in retrieving the weaker associates (as evidenced by these metrics) resulted in a memorial benefit over time, with a final cued recall test and a delay of 3-5 days. In the test phase analysis, there was evidence for an interaction between accuracy at each test phase and association strength. The pattern showed that less retention was obtained for the weaker associates from initial to final test than for the stronger associates, consistent with the results from experiment 2. Finally, for the testing effect analysis, both a reliable main effect of practice task and association strength were produced, however again no evidence for an interaction was found. Furthermore, evidence for no interaction was found based on the bayesian evidence from the conditional analysis. Taken together, the results suggest, that although there seemed to be more elaborate processing and increased difficulty present in the

learning of the weaker associate pairs, this did not translate into a larger testing effect for these items.

2.5 General Discussion: Experiments 1-3

Across three experiments the results from Carpenter (2009) were replicated (experiment 1) and further explored (experiments 2 & 3) in relation to whether they could extend to a well established testing effect phenomenon, the test-delay interaction. The test-delay interaction is the finding that as the delay to the final test increases, the size of the testing effect also increases (Adesope et al., 2017; Rowland, 2014).

Across all three experiments there was evidence that weaker associates took longer to retrieve during the practice task and were rated as less associated than the stronger associates. This is consistent with the ideas of the elaborate retrieval hypothesis, that a more elaborate search is being conducted when the links between items are weaker. However, rather unexpectedly, for both the elaborative retrieval hypothesis and the testing difficulty hypothesis, there was a lack of evidence to suggest that increased difficultly during practice retrieval for the weaker associates, as evidenced by the time to respond during the initial test, resulted in a testing effect benefit over time. In addition, the main testing effect analyses across the three experiments found no supporting evidence in favour of the elaborate retrieval hypothesis (testing effect analysis and conditional analysis, experiments 2 and 3), instead some evidence for the null hypothesis was found (testing effect analysis and conditional analysis, experiment 1).

The conditional analysis conducted here in replication, has been criticised elsewhere for introducing item selection effects (Carpenter, Pashler, Wixted, & Vul, 2008; Kornell et al., 2011). In this way easier to retrieve items are thought to show reduced forgetting (retrieval advantage) at final test as a function of stronger item memory strength at the time of initial retrieval. Future experiments in this series will not make a conditional analysis assessment, as no demonstrable utility has been shown in furthering understanding in the domain of interest. Furthermore, the contribution of item encoding is of interest in this thesis alongside the direct effects of testing. Therefore, any

analysis that could potentially mask the absolute magnitude of the testing effect with the different learning materials employed herein could negatively impact the learning gained from this series.

The ERH suggests that a longer search should result in a greater benefit of retrieval practice. The time phase analysis in experiment 1 was consistent with Carpenter's hypothesis and previous results, demonstrating that at the final free recall task the weaker associates no longer showed a deficit in comparison to the stronger associates, that was seen on the initial practice test. Therefore, the time phase analysis from experiment 1 alone is consistent with the elaborate retrieval hypothesis.

However, there are a couple of issues with this supporting evidence. Firstly, although the time phase analysis in experiment 1 does show a benefit in retention for weak associates come final test, we cannot reasonably equate this to a testing effect benefit for weak associates due to the lack of restudy comparison for this data. Therefore, this pattern could just be a function of any form of memory practice task. In addition, although we are talking about a retention benefit, due to the change in test type between initial and final test in experiment 1, we do not have a consistent measurement for retention. Rather, the time phase analysis of experiments 2 and 3 more accurately depict retention rates, based on the matched retrieval tasks between initial and final test. In these cases the evidence was in favour of a strong associate benefit at the final test. The time phase results in experiments 2 and 3 stand in conflict with the predictions of the elaborate retrieval hypothesis.

These points combine to question whether Carpenter's original experiments contained particular features that revealed the compelling results. One such feature could have been the change in test types from an initial cued recall test to a final free recall test. Although, this idea has not been explored in relation to the mechanisms of the testing effect. As evidenced here, features of a change in test type could influence how likely it is to find a testing effect. It could also influence the nature of the conclusions drawn based on utilising a particular test type. Therefore, it might be pertinent for future research to explore any main findings in relation to a variety of test types and changes

between initial and final test types. Elsewhere such changes have previously resulted in changes to the pattern of results associated with the testing effect (Hinze & Wiley, 2011).

A second feature of the original compelling results, could have been Carpenter's inclusion of short lists of word pairs in experiment 1, where the stronger evidence was found, rather than the long lists of word pairs that were utilised in experiment 2 and in this series. Here it is reported that long lists of word pairs do not show a strength benefit through testing via cued recall, although further work should look to explore whether shorter lists of word pairs tested via free recall provide the same outcome. However, any benefit of free recall as reported here and previously by Carpenter must also take into account the low accuracy rates that were seen. It is possible that delivered in smaller doses and perhaps in concert with cued recall methods this could be an effective way to boost retrieval for less familiar or more difficult to retrieve items, where testing effects can sometimes be evasive (Van Gog & Sweller, 2015).

The results reported here across three experiments, suggest that Carpenter's results from 2009 should not be taken to provide broad support for the elaborate retrieval hypothesis of the testing effect. In relation to this I suggest that the metrics for difficulty as reported here by longer response times for the weak associate pairs, are not useful for assessing the mechanisms of the testing effect, yet could still indicate the presence of elaborate processing. It is hard to understand, given the previous work in this area, how difficult to retrieve items did not seem to benefit from testing here in comparison to restudy practice. However, as already outlined in chapter one the concept of difficulty is also not clearly defined, beyond a greater amount of spacing between item presentation (Carpenter & Yeung, 2017), or test phases (Pyc & Rawson, 2009) or the number of cues being provided (Carpenter & DeLosh, 2006). In any case there is also a condition by which increasing difficulty does not lead to a benefit (Van Gog & Sweller, 2015) and issues surrounding this work appear to be poorly understood. In addition, as evidence has already suggested (Carpenter & DeLosh, 2006; Cho et al., 2017; Rowland, 2014), it is possible that the nature of the retrieval task and not the study items

is where difficulty and possibly elaboration counts. Therefore, understanding where difficulty is most influential during testing will further develop our understanding of the mechanisms relevant to testing.

The three experiments detailed here were designed to explore the validity of the original findings relating to the elaborate retrieval hypothesis and more broadly the contribution of meaningful processing in the study materials to the testing effect, which to date has only been tentatively captured. The results reported here struggle to support both the original hypothesis and through this the idea that meaningful processing more broadly contributes to the testing effect.

Chapter 3

Meaningful processing via mediation and structural coherence

The current chapter explores two different aspects of meaningful processing in relation to the study materials. Firstly, experiment 4 examines the extent to which ease of meaningful processing might be driving the testing effect, based on the mediator effectiveness hypothesis. In experiments 5 and 6 the idea of meaningful processing will be explored through the structural coherence of the items being studied, based on work in the retrieval-induced forgetting literature.

3.1 Introduction

The results from chapter one do not provide compelling evidence for the elaborate retrieval hypothesis or that meaningful processing contributes to the testing effect. However, a broader exploration of the role of meaningful processing in the study materials is now required to assess how comprehensive the results of chapter one are. To this end, there are two further angles that will be explored in this chapter. The first, explored in experiment 4 is the mediator effectiveness hypothesis and the second, explored in experiments 5 and 6 is the structural coherence of the study materials.

Developed in complement to the elaborate retrieval hypothesis, the mediator effectiveness hypothesis relies on similar cognitive processes, but makes slightly different predictions about when retrieval practice will be most beneficial. The mediator effectiveness hypothesis, was first developed by Pyc and Rawson (2010) and subsequently supported by further empirical work (Carpenter, 2011; Carpenter & Yeung, 2017; Rawson et al., 2015). It suggests that retrieval practice makes better use of mediating information available during the retrieval practice task than the restudy practice task.

This use of mediating information during the retrieval practice task results in stronger links between cue and target than via restudy practice. These stronger links occur because additional retrieval routes are activated when mediating information is utilised during retrieved practice. These additional retrieval routes in turn mean that final test accuracy is improved for the retrieval practice items when compared to the restudy practice items. For example, the word pairs in experiments 1-3 consisted of a cue and target where the target word in each case was the highest associated word to the cue. This meant that there was no mediating information linking the cue to the target, because the target was the highest associated link. There would have been mediating information between cue and target if the target was not the strongest associate to the cue. In this way the mediator, could have served as an additional retrieval route from the cue to the target.

We currently have no evidence from a direct manipulation of the accessibility of mediating information during the retrieval task. However, based on the previous work, it is reasonable to suggest that through a direct manipulation of the materials we could assess whether study materials that activate more mediating information benefit more from retrieval practice than study materials that do not. Ideas consistent with this suggest that testing is able to promote relational processing between items (Congleton & Rajaram, 2012; Zaromb & Roediger, 2010). If this is the case, then we would expect materials where mediating information is available to assist retrieval relative to restudy more than where no or little mediating information is available. The original work in this area (Pyc & Rawson, 2010) found that when participants learned Swahili-English translation pairs with the help of self-generated mediators, test practice demonstrated a larger benefit over restudy practice. During the encoding phase, participants were instructed to provide some additional mediating information that would help them remember the items, which participants were later prompted to retrieve each time they restudied the pair. For example, if the pair to be learned was *wingu-cloud*, the participants were encouraged to create a keyword that looked or sounded like the Swahili cue and was semantically similar to the English word, in this case the keyword generated

might be *wing*. Participants that generated mediators in combination with retrieving the target during the practice task performed better at final test than participants that had generated the mediators during restudy only. This suggested to authors that learning word pairs with the help of mediators was more beneficial with test practice, due to test practice boosting the utility of mediators which assisted future retrieval.

There is a long history in memory research of studies demonstrating that some form of mediation or relational processing is useful for retrieval (J. R. Anderson & Reder, 1979; Craik, 2002; Einstein et al., 1990). Understanding how these findings apply to the testing effect is important to future theoretical and practical developments in this area. As outlined above there is currently relatively little evidence that has directly explored this concept. Therefore, if mediation is key to retrieval practice effects associated with the testing effect, a more direct manipulation of this idea, by providing both helpful mediators and less helpful mediators, should reveal results in support of the mediator effectiveness hypothesis. Furthermore, this is a second way to assess whether meaningful processing contributes to the testing effect.

Following experiment 4, the focus of this chapter turns to a different area of the literature that has indicated evidence that meaningful processing in the study materials contributes to the testing effect. This is the idea that the structural coherence that the study materials are presented in impacts the testing effect. This concept is feasible if we consider the support from Rowland's (2014) meta-analysis. This showed that organised materials lead to larger testing effects than less organised materials. For example, prose and paired associates benefit more from testing than single words or unrelated items. However, it is difficult to note what quality about these information structures leads to the differences observed.

Early reports suggested that the organisation of materials was important to how much benefit was gained through testing (Gates, 1922). However, not much work has been done in this area and results are contradictory (Chan, 2009; de Jonge et al., 2015). One suggestion could be that increased organisation in text helps the reader to build a mental model based on the progression of semantically linked ideas (Foltz,

Kintsch, & Landauer, 1998), which might be lacking in less organised text structures. These texts are more likely to benefit from cues that draw on a combination of semantic (Carpenter, 2011; Pyc & Rawson, 2010) and temporal properties during retrieval practice (Karpicke et al., 2014) than when items are less organised.

In line with these ideas Chan (2009) found that when items were coherently organised, in an easy to comprehend text passage then this information was more easily retrieved on the final test than when items were less coherently organised or randomly organised, when compared to a no testing control. These results were thought to reflect a testing advantage for the coherent items. However there was no matched restudy control task, leaving the possibility that the results reflected lower recall or increased inhibition for the low coherent materials, relative to a no study control (see Shimmerlik, 1978).

Contrary to these results, de Jonge et al. (2015) found that when items were presented in a less coherent manner, whereby the items were presented in a scrambled sentence order, then testing was more advantageous than when items were presented in a coherent form. However, results came from two separate experiments rather than a direct manipulation in one experiment and the time spent studying each item of the low coherence text was longer than the higher coherence text although the total time was matched, suggesting another possible explanation for the findings.

As the results from Chan (2009) do not have the presence of an adequate restudy control and the results of de Jonge et al. (2015) were not manipulated within the same experiment, the findings here still require further exploration.

Levels of coherence in texts can have a direct impact on comprehension (O'Reilly & McNamara, 2007) and is therefore an important line of enquiry for research into the application of the testing effect. How textual coherence influences the testing effect was explored in experiments 5 and 6, under stricter controls than previously observed. In experiment 5 the text materials reflected work that had been previously done, replicating more readily applied educational materials (Chan, 2009; de Jonge et al., 2015). In experiment 6 there was a further effort to employ a more extreme manipulation of

text coherence. By assessing the differences in the testing effect based on textual coherence in these two experiments, I am further able to assess a different element of meaningful processing in relation to the testing effect.

3.2 Experiment 4

Experiment 4 aimed to directly address the claims of the mediator effectiveness hypothesis, as to date there has not been a direct manipulation of mediation in the literature. Some of the results of the previous work in this area has relied on repeat cycles of restudy and test practice (Pyc & Rawson, 2010), which we know is likely to inflate any benefits associated with testing (Eglington & Kang, 2018; Kang et al., 2007; Rowland, 2014). For example, as there are both mediating effects of retrieval on both subsequent encoding and feedback on subsequent retrieval, a direct manipulation of the contribution of mediators to the testing effect based on the direct effects of testing, without the addition of feedback, is required to assess the utility of the mediator effectiveness hypothesis in relation to the testing effect.

Previous work in this area has assessed the contribution of the existing associative networks of word pairs (Carpenter, 2011; Carpenter & Yeung, 2017; Rawson et al., 2015), made use of repeat cycles of testing (Carpenter & Yeung, 2017; Rawson et al., 2015) and non-standard assessments of the testing effect (Carpenter, 2011). For example, based on whether mediators are recognised as having been part of the study set. Some recent work has shown that the effect of semantic mediating information might be smaller than originally thought (Coppens, Verkoeijen, Bouwmeester, & Rikers, 2016). Therefore, there is still much work to do in this area in establishing how mediators contribute to the phenomenon of the testing effect.

Experiment 4 will assess whether providing participants with information that can be utilised as semantic mediators benefits the testing effect to a greater extent than information that participants are less able to utilise as semantic mediators. Experiment 4 will utilise similar materials to Pyc and Rawson (2010), by providing participants with word pairs to learn which are difficult to link together without the help of additional, *me-*

diating, information. The direct manipulation here is whether this mediating information is helpful to forming semantic links between the paired words or not.

Due to the nature of the constructed materials, there are some additional points to be aware of in the design of this experiment that deviates from the previous chapter. As trial stimuli were likely to be unfamiliar to participants, it was necessary to achieve sufficient familiarity during the initial encoding phase. This was to ensure that a good level of retrieval was achieved, which is important to be able to detect a testing effect (Rowland, 2014). For this a deep encoding task was given, whereby participants were asked to try to use the mediating information provided to create a vivid image in memory for the word pair. Experiment 4 included two test conditions, one that included feedback and one that did not. A feedback condition was added due to the increased difficulty of the task. Feedback use is advised for occasions when initial performance might not be particularly high (Kornell et al., 2011; Rowland, 2014). In addition, due to the fact that Pyc and Rawson (2010) utilised repeat restudy-test trials, including a feedback condition was necessary to see if this alone was capable of influencing the pattern of results relative to the test only condition. Here a feedback condition was included that offered a restudy opportunity but not a repeat retrieval opportunity, thereby, reducing the additive properties that could have influenced the results found previously by Pyc and Rawson (2010).

Experiment 4 also included a further study condition. As the series in chapter two failed to find supporting evidence for the elaborate retrieval hypothesis, here it was possible to include a different assessment of the ERH. Previous work has shown both that using elaboration techniques can be useful for learning with novel materials (Willoughby et al., 1994) and that these techniques can be equivalent to retrieval practice efforts (McDaniel et al., 2009). An elaborate restudy task was added to the practice conditions, to again help to boost memory rates in the final test with these unfamiliar materials and to explore previous work on meaningful processing through elaboration.

Furthermore, in experiment 4 it was necessary to utilise an immediate final test design, such that the final test occurred within the same study session. Studies have

previously shown a reliable albeit smaller testing effect when the final test is administered within the same session (Rowland, 2014). The immediate test was necessary as the learning materials for experiment 4 were less familiar and phonologically more complex than those in chapter 2 making them harder to retain over time (Hulme, Maughan, & Brown, 1991). In addition as already specified above, experiment 4 was interested in exploring the mediator effectiveness hypothesis when repeated restudy-test opportunities were not part of the design. As a repeated restudy practice opportunities increase accuracy for items (Roediger & Karpicke, 2006b; Rowland, 2014; Wheeler et al., 2003), absence of this element in experiment 4 meant an immediate final test was the preferable.

A prediction is made based on the mediator effectiveness hypothesis, that word pairs accompanied by helpful mediators will utilise these mediators more and consequently benefit to a greater extent from retrieval practice, in comparison to word pairs accompanied by less helpful mediators.

3.2.1 Methods

Participants and Design

Participants were 120 students at the University of Plymouth, aged between 18 and 29 (M = 20.83, SD = 2.19), 74.4% female. Participants took part in the study for course credit or were paid for their time at £8 per hour, £2 each 15 minutes.

The study utilised a 4 x 2 mixed design, with practice task (restudy, elaborate restudy, test, test with feedback) as a between-subjects factor and accompanying definition language (English, Swahili) as a within-subjects factor. The English definition language represented a more helpful mediator manipulation and the Swahili definition language represented the less helpful mediator manipulation. The final test in experiment 4 was administered immediately (after 5 minutes) as this has been demonstrated to be ample time for a testing effect to arise (Rowland, 2014).

Sample Size Calculation

The sample size was calculated based on the effect sizes found in the Pyc and Rawson (2010) paper, for which the current materials were quite a close match. Pyc and Rawson (2010) found a large effect (conservative estimate > 1) in the cue only final test condition, which would represent the strong mediator condition in the current experiment. We expected, in the absence of the repeated retrieval design and with a 5 minute delay to final test we would gain a smaller effect. We anticipated we might find a small to medium interaction effect (f = 0.16) with these materials. The sample was calculated, based on having four between-subjects groups and two measures for each. A total sample size of 112 was required based on the G*Power (Faul et al., 2009) calculation, with power of 0.80, this was rounded up to 120, making 30 per group rather than 28.

Materials

Each participant saw 50 word pairs, consisting of 2 practice word pairs and 48 word pairs split into three lists of 16 word pairs. Each pair consisted of a rare adjective cue word and common noun target, for example *Capricious-Pigeon*. Adjectives were rare English words that ranked at under 11 per million or fewer on the MRC psycholinguistics database (Coltheart, 1981). Nouns were words ranked at between 20 and 100 per million and were also taken from the MRC psycholinguistics database. Each word pair was accompanied by a definition for the rare adjective during the familiarity task and the initial study phase. Definitions for the rare English words were taken from Collins online dictionary and then shortened to between two and five key words for brevity and consistency that captured the word's meaning. For example, *capricious* defined as, "Something that is capricious often changes unexpectedly" became "unpredictable, impulsive". This was to enable participants to utilise the definition without having to spend too much time reading it. Adjectives and nouns were paired up randomly, each participant saw the same adjective-noun pairing, although in a new random order for each participant and during each presentation phase. The shortened definitions were

then translated into Swahili via Google translate. The length of the number of characters for the definitions in each language were not significantly different from each other ($t(47) = 1.70$, $p = .10$, $d = 0.25$, $BF_{10} = 0.60$), with the English definitions ($M = 31.83$, $SD = 12.54$) including a similar number of characters to the Swahili definitions ($M = 30.52$, $SD = 14.41$). The full list of word pairs for experiment 4 can be found in appendix B.1.

Procedure

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. The experimental session took 45 minutes. Participants were tested in groups of up to six people. For the duration of the experimental session, participants were sat at a partitioned desk with their own PC. Participants wore headphones throughout the task. Besides the filler task, all elements of the experimental task were presented using E-Prime 3.0 software (Psychology Software Tools, Inc, Pittsburgh, PA, 2016). Each list of sixteen cue-target pairs followed the same procedure; a familiarity task, a study phase, a practice phase and a test phase. Participants were instructed that they would not know what type of practice trials they would have until they entered the practice phase for each list. However, all participants completed the same practice task for each of the lists, assigned to one of four conditions; restudy, elaborate restudy, test, or test with feedback. Participants received two practice items prior to starting the first list, which included an example of the familiarity task, a study trial and test trial, including an example of the word pair accompanied by an English definition or Swahili definition. For the participants in the test with feedback condition their practice trials included a test with feedback trial. Each rare English word was assigned a definition in either English or Swahili for the duration of the study. This definition was presented at each phase that it would be included (familiarity task, study phase and for the elaborate restudy condition the practice phase also). Once participants had completed one list of sixteen word pairs through the four different phases of the task, they repeated the process for a further two lists of sixteen word pairs, making

up 48 word pairs in total.

The materials were designed so that participants would not easily be able to form a connection between the cue and target without the assistance of the mediating information. As such, the familiarity task was designed to verify that participants did not have considerable knowledge of the cue words and to ensure that participants would achieve a moderate degree of recall through an additional exposure to the cue words and the definitions. Participants saw each rare adjective in the centre of the screen for 3 seconds, followed by the question "How familiar is this word?" Participants were instructed to make a response from 1 to 5 on the scale: 1 = Never seen or heard the word before, 2 = Seen or heard the word but unsure of its meaning, 3 = Some understanding of the word, 4 = Fairly confident I know what the word means, 5 = Confident using this word in a sentence. Once participants had made a response, the adjective remained on screen but now with its definition beneath it. The definition was presented in either English or Swahili and participants were asked "How confident are you that you will remember the definition?" Participants were instructed to make a response between 1 and 5, 1 being not confident and 5 being very confident. The definition was presented during the familiarity task to ensure that participants in all conditions saw the definition more than once before entering a practice test phase. It also ensured that during the study phase the Swahili items in particular were not completely novel to participants. Once all 16 items in the list had been rated for familiarity and each definition presented, participants moved on to the study phase.

In the study phase, participants were presented with each of the 16 word pair items in the current list in a new random order. Participants saw the adjective cue on the left of the centre of the screen and the noun target on the right of the centre of the screen. Beneath the word pair was the same definition as previously presented with the adjective cue in the familiarity task, again either in English or Swahili (same as viewed in familiarity task). The three items remained on screen for 8 seconds, with a prompt at the top of the screen for participants to use the definition to create a vivid memory for the word pair. After 8 seconds, participants were asked "How useful is the

definition in creating a vivid memory of the word pair?" Participants were instructed to make a response between 1 and 5, 1 being not useful and 5 being very useful. Once the participant had made a response the program moved on to present the next word pair. Each study trial was followed by a 500ms intertrial interval. Once participants had studied each of the 16 word pairs they moved on to the practice phase for the list.

For the practice phase, stimuli were presented in a new random order and participants completed a practice task depending on the condition they had been assigned to. For participants in the elaborate restudy condition this was the same as the study phase, whereby participants were again asked to use the definition to create a vivid memory for the word pair and rate the usefulness of the definition in achieving this. In the restudy condition, participants were only shown the word pairs again, without any definition present or making a rating. For participants in the test conditions, each cue word was presented with the question "Can you remember the target word?" Participants were instructed to press the space bar and type in the word they recalled and press enter. If participants could not recall the target word they were instructed to type in "no" and press enter. For participants in the feedback condition, once they had entered their response the correct cue and target words were displayed again on screen for 3 seconds. Each trial was followed by a 500ms intertrial interval.

Following the practice phase participants completed a 60 second filler task, in which they worked on a number search (with pen and paper). A numerical filler task was chosen for the break, to minimise the likelihood of participants rehearsing something associated to the studied materials. Following the task participants moved on to the final test phase.

For the final test phase, all participants completed the same task. Each of the 16 cue words were presented in a new random order, the final test phase was in the same format as the practice test, whereby for each cue participants were asked "Can you recall the target word?" Participants were asked to respond either by typing in the word they had recalled or by typing "no" if they could not recall the target word. Each trial was followed by an intertrial interval of 500ms.

Following the final test phase participants completed the same procedure for the remaining two lists, all participants completed the same learning task for each of the three lists. Following which they were debriefed and dismissed.

3.2.2 Results

All analyses were conducted in JASP (JASP Team, 2020). All frequentist analyses where appropriate are given with the results of bayesian equivalent analyses. Descriptive statistics for the main effects of interest are given in table 3.2.

Coding Responses

Items were coded blind to condition. Plurals incorrectly present or absent, obvious spelling mistakes and two letter changes to make up correct words (but not another word) were coded as correct. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible.

Ratings and Response Times

Initial analyses were conducted on the numerous ratings made throughout the task. Table 3.1 gives the descriptive statistics for the various ratings collected in experiment 4.

The average familiarity rating for the rare English adjectives was 1.75 out of 5, suggesting that individuals were not familiar with the cue words. This rating was not different between the two definition conditions, English (M = 1.75, SD = 0.56) and Swahili (M = 1.75, SD = 0.57) ($t(119) = 0.03$, $p = .98$, $d = 0.003$, $BF_{10} = 0.10$), this was to be expected as ratings were made prior to exposure to the language definition manipulation.

Participants then rated the helpfulness of the definition during the familiarity task, a 2 x 4 mixed ANOVA was conducted with definition language as a within-subjects factor and practice task as a between-subjects factor. There was a main effect of definition language on definition rating, with the English definition (M = 2.70, SD = 0.71) rated as more helpful than the Swahili definition (M = 1.20, SD = 0.27), $F(1,116) = 699.88$,

$p < .001$, $\eta_p^2 = 0.86, BF_{10} > 150$ (1.771e+47), but not between the different practice tasks, $F(3,116) = 1.25$, $p = .30$, $\eta_p^2 = 0.03, BF_{10} = 0.19$, and no interaction was present between definition language and practice task, $F(3,116) = 1.47$, $p = .23$, $\eta_p^2 = 0.01, BF_{10} = 0.20$.

In the study phase participants were asked to rate the vividness of the word pairing, again a 2 x 4 mixed ANOVA was computed, with definition language as a within-subjects factor and practice task as a between-subjects factor. There was a main effect of definition language with English definition word pairs rated as being more memorable ($M = 2.62$, $SD = 0.74$) than the Swahili definition word pairs ($M = 1.40$, $SD = 0.50$), $F(3,116) = 353.33$, $p < .001$, $\eta_p^2 = .75, BF_{10} > 150$ (3.200e+38). Vividness ratings were equivalent across practice tasks, $F(3,116) = 1.63$, $p = .19$, $\eta_p^2 = .04, BF_{10} = 0.09$. There was no interaction between practice task and definition language ($F(3,116) = 0.97$, $p = .41$, $\eta_p^2 = .01, BF_{10} = 0.14$).

Finally, the response times to correct and all responses between the two types of definition language trial during the retrieval practice task were compared. Results showed that English definition test practice trials resulted in longer response times than the Swahili definition test practice trials, both for correct responses (English, *Mdn*=3176, Swahili, *Mdn*=2661) and across all responses (English, *Mdn*=3859, Swahili, *Mdn*=3534). Wilcoxon signed-rank tests were conducted due to violation of normality (Correct responses, T=1121, $p<.001$, r=.41.; All responses, T=1417, $p<.001$, r=.55. Bayes factor not calculated due to the violation of normality). This finding although somewhat counterintuitive based on the desirable difficulties argument, whereby more difficult items should result in longer response times. This could instead reflect a greater search during test practice for the English definition pairs, as more mediating information was available, as opposed to the Swahili definition pairs, where less mediating information was available, which could have led to early termination of the memory search.

Table 3.1
Familiarity, Helpfulness and Vividness Ratings during Familiarity and Study Tasks and Response Times during Retrieval Practice in Experiment 4 as a Function of Definition Language

Measure	English definition	Swahili definition
Familiarity rating (familiarity rask)	1.75 (0.56)	1.75 (0.57)
Helpfulness definition rating (familiarity task)	2.69 (0.72)	1.16 (0.27)
Vividness of memory rating (study task)	2.62 (0.74)	1.43 (0.50)
Response times (RTs) – correct answers (s)	3.76 (2.64)	2.83 (1.22)
Response times (RTs) – all answers (s)	4.62 (2.29)	3.97 (1.96)

Note. Mean values represented for ratings and RTs, RTs given in seconds, SDs given in parentheses.

Main Analyses

Initial test performance. Performance on the initial test was assessed by a 2 (practice task; test, test with feedback) x 2 (definition language; English, Swahili) mixed ANOVA, with practice task as a between-subjects factor and definition language as a within-subjects factor. Results showed a main effect of definition language, with Swahili definition pairs (*M* = .20, *SD* = 0.17) leading to poorer cued recall than the English definition pairs (*M* = .30, *SD* = 0.16), $F(1,58) = 24.93$, $p < .001$, $\eta_p^2 = 0.30, BF_{10} > 150$ (1685.77). There was no main effect of practice task during initial test performance, as the test (*M* = .26, *SD* = 0.18) and the test with feedback group (*M* = .25, *SD* = 0.16) were evenly matched during the initial retrieval task, $F(1,58) = 0.50$, $p = .48$, $\eta_p^2 = 0.10, BF_{10} = 0.32$. There was however evidence for an interaction between definition language and practice task test condition $F(1,58) = 5.32$, $p = .03$, $\eta_p^2 = 0.08, BF_{10} = 2.21$, although results of the bayes factor are inconclusive ($0.33 < BF_{10} < 3$). Follow-up paired samples t-tests by practice condition revealed strong evidence for a difference between the English (*M* = .31, *SD* = 0.19) and Swahili (*M* = .16, *SD* = 0.15) definition language for the test with feedback condition, $t(29) = 5.04$, $p < .001$, $d = 0.92$, $BF_{10} = 995$. But no clear evidence for a difference between the English (*M* = .29, *SD* = 0.14) and Swahili (*M* = .23, *SD* = 0.18) definition language for the test only condition, $t(29) =$

Table 3.2
Initial and Final Test Accuracy in Experiment 4 as a Function of Practice Task and Definition Language

	Initial test		Avg	Final test		Avg
Practice task	English	Swahili	IT	English	Swahili	FT
Restudy	n/a	n/a	n/a	.46 (0.27)	.44 (0.27)	.45 (0.26)
Elab restudy	n/a	n/a	n/a	.49 (0.16)	.36 (0.18)	.42 (0.15)
Test	.29 (0.14)	.23 (0.18)	.26 (0.14)	.29 (0.15)	.22 (0.17)	.25 (0.14)
Test with FB	.31 (0.19)	.16 (0.15)	.23 (0.15)	.48 (0.23)	.35 (0.20)	.42 (0.19)

Note. The values represents mean percentages of target words recalled, SDs given in parentheses.

1.95, p = .06, d = 0.36, BF_{10} = 1.02.

Final test performance. Final test performance was assessed in a 4 x 2 mixed ANOVA, with definition language (English, Swahili) as a within-subjects factor and practice task (restudy, elaborate restudy, test and test with feedback) as a between-subjects factor. A main effect of practice task was found, $F(3,116) = 6.50$, $p < .001$, $\eta_p^2 = 0.14, BF_{10} = 74.37$, suggesting that final test performance was not uniform across all groups. Follow up analyses showed that, *test with feedback* performed comparably to *elaborate restudy* ($t(58) = 0.25$, p = .81, d = 0.05, BF_{10} = 0.27) and *restudy* ($t(58)$ = 0.64, p = .52, d = 0.14, BF_{10} = 0.30), but better than *test* ($t(58) = 4.39$, $p < .01$, d = 1.13, BF_{10} = 64.34). *Elaborate restudy* and *restudy* also performed better than test (without feedback) on final test performance ($t(58) = 5.25$, $p < .001$, d = 1.36, BF_{10} = 692.3 and $t(58) = 4.68$, $p < .01$, d = 1.21, BF_{10} = 45.37 respectively). The *elaborate restudy* group performed comparably to the *restudy* group on final test performance, $t(58) = 0.48$, p = .64, d = 0.11, BF_{10} = 0.28.

In addition, a main effect of definition language was found, showing that final test performance was superior for the English definition trials (M = .43, SD = 0.22), in comparison to the Swahili definition trials (M = .34, SD = 0.22), $F(1, 116) = 36.8$, $p <$ .001, $\eta_p^2 = 0.24, BF_{10} > 150$ (323386.2).

In line with our a priori hypothesis, an interaction was indicated between practice task and definition language, based on the frequentist results, although not the bayes factor which was inconclusive ($F(3, 116) = 2.92$, $p = .04$, $\eta_p^2 = .07, BF_{10} = 1.13$). Follow-up tests revealed this was not in the predicted direction. Paired samples t-tests for each practice task condition revealed that the targets from the English definition pairs were retrieved more accurately than the targets in the Swahili definition pairs for the elaborate restudy condition ($t(29) = 4.22$, $p < .001$, $d = 0.77$, BF_{10}=128.10), the test condition ($t(29) = 2.72$, $p = .01$, $d = 0.50$, BF_{10}=4.17) and the test with feedback condition ($t(29) = 4.02$, $p < .001$, $d = 0.73$, BF_{10}=77.77), but not the restudy condition ($t(29) = 0.90$, $p = .38$, $d = 0.16$, BF_{10}=0.28). Figure 3.1 depicts this result.

One final 2 x 2 ANOVA compared final test performance for the language definitions (English, Swahili) and the type of test practice task (test only, test with feedback). Results revealed a main effect of practice task with test with feedback ($M = .42$, $SD = 0.19$) showing higher accuracy for targets than the test only condition ($M = .25$, $SD = 0.14$), $F(1, 58) = 14$, $p < .001$, $\eta_p^2 = 0.19, BF_{10} = 68.82$. There was a main effect of definition language, with English definition targets ($M = .38$, $SD = 0.21$) being recalled more accurately than Swahili definition targets ($M = .28$, $SD = 0.20$), $F(1, 58) = 23.29$, $p < .001$, $\eta_p^2 = 0.28, BF_{10} > 150$ (1353.81), but no evidence for an interaction between these two factors, $F(1, 58) = 1.81$, $p = .18$, $\eta_p^2 = 0.02, BF_{10} = 0.53$.

3.2.3 Discussion

The results from experiment four demonstrate a successful operationalisation of the mediator helpfulness manipulation, namely as English definition word pairs were recalled more often than Swahili definition word pairs. Prior to the definition pairings, participants rated the unusual English words, as equally unfamiliar for those assigned to the English and Swahili definition conditions. Participants also rated the English definition as being a clearer link to the English adjective and further rated the English definition as better able to help them form a more vivid memory for the word pair. These ratings corresponded to increased accuracy during recall, which on average

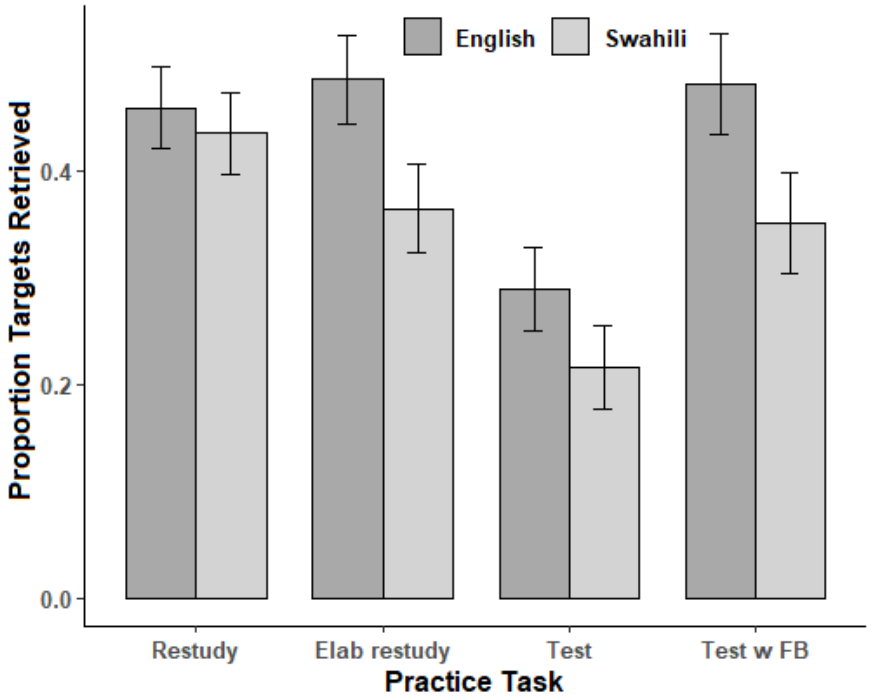

Figure 3.1. Mean target retrieval at final test as a function of practice task and definition language in experiment 4. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

took longer to trigger when items were paired with an English definition than a Swahili definition. This is consistent with the possibility that the English definition is being used more effectively as a mediator.

One point to note with the current findings is that no testing effect was found here, instead results showed the pure testing condition performed more poorly than all other conditions on these materials. These results could be consistent with a complexity account, whereby items that require the integration of multiple elements to comprehend can lead to a negative or non-existent testing effect (Van Gog & Sweller, 2015). This has been found for example when study materials detail component parts that are required to solve a problem on the final test. This could be viewed as conceptually similar to the current study materials on the Swahili trials. The fact that the restudy condition showed the best recall for the Swahili definition pairs, demonstrated that integrating the information was not an easy task. In fact, it could suggest that when items need to be integrated, providing non-helpful mediators is not as useful as rote memory. Previous

work has shown that for restudy tasks participants rely on rote memorisations more so than for retrieval tasks when incidental associations can be made between items (Cho & Powers, 2019). This study found both better verbatim and conceptual learning for word definitions for Chinese characters in the retrieval practice group than the restudy group. The items to be learned were novel but making associations between items was helpful for a final conceptual test. For example, Chinese character similarities denoted similar meanings. It should be noted that in this study there were two restudy-test blocks, where correct answer feedback was given after a retrieval attempt. As making associations between items would not have been possible in the retrieval practice condition for the Swahili items in experiment 4, it is possible that rote memorisation and high difficulty items could explain the restudy benefit found for Swahili restudy items here.

The lack of testing effect is also consistent with the bifurcated distribution account (Kornell et al., 2011), whereby because the materials were not adequately recalled during practice they did not benefit from retrieval in comparison to restudy. However, testing effects have been found in the presence of low accuracy (Hinze & Wiley, 2011), albeit when the delay to final test was greater. One way to compensate for the low retrieval rates would be to replicate the helpfulness manipulation with easier to integrate materials or with a revised design that allows for finding a testing effect. For example, this could be achieved by increasing the delay to final test, or with an extended encoding period to increase initial retrieval levels. [1]

Despite the fact that no testing effect was found here, as the meaningful processing manipulation was present, we could expect to find evidence for the mediator effectiveness hypothesis. If mediation was made more effective in combination with retrieval practice, at final test we would expect a testing effect on the English definition trials relative to the Swahili definition trials. Based on Pyc and Rawson's results, this would be more pronounced in the test with feedback condition than the test only condition.

[1] In attempting to address this issue, two pilot studies were conducted, with varying delay periods, to assess the possibility of completing a delay to the final test, to encourage a testing effect with these materials. However, it was not viable to run these studies in full due to accuracy rates being at floor.

Relevant to this and the results of Pyc and Rawson (2010) is the difference between the test only condition and the test with feedback condition. There was no interaction shown in relation to the mediator manipulation, however the test with feedback condition showed greater overall accuracy at final test. Furthermore, in comparing the restudy condition with the test with feedback condition, there was no statistical difference on the English definition trials. However, there was a numerical difference favouring the test with feedback condition, possibly suggesting that with repeated cycles as seen in the Pyc and Rawson (2010) study, this difference could become significant. Based on the current results, the direct effects of mediation on retrieval appear to be non-existent with difficult to learn items, but is likely to be at best small with items that are not difficult to learn and have pre-existing associations (Coppens et al., 2016; Hausman & Rhodes, 2018).

Relevant to this point is the finding that a restudy opportunity, in both the case of the pure restudy condition and the test with feedback condition, was beneficial over the pure test condition. The results also showed that the restudy condition and the test with feedback condition performed comparably to the elaborate restudy condition, whereby participants were given two opportunities to view the definitions and try to create a vivid memory for the the word pair for later retrieval. This suggests that retrieval practice is not always superior to restudy or elaborate restudy, as has been reported previously (Blunt & Karpicke, 2014; Coane, 2013; Karpicke & Blunt, 2011; Lechuga, Ortega-Tudela, & Gómez-Ariza, 2015; McDaniel et al., 2009). Although interestingly, McDaniel et al. (2009) found that an elaborate study task (note-taking) performed at similar levels to a retrieval practice condition (read-recite-review) when the study materials were longer and more complex in experiment 2 than in experiment 1. Recent work has found that items more complex (Roelle & Berthold, 2017) and lower in cohesion (Roelle & Nückles, 2019) benefit less from testing. Whether the pattern of results in experiment 4 can be attributed to low accuracy (Kornell et al., 2011), or to the level of elemental interactivity in the study materials (Van Gog & Sweller, 2015) is not clear.

Although there was no testing effect to report here, the current results are consistent

with findings from the experiments given in chapter two. There is no evidence with the design and materials of experiment 4 that testing interacts with either the helpfulness of the mediators or the ability for participants to elaborate during the practice task as was found in chapter two. What is possible however, is that particular properties of the materials might have led to not finding a testing effect. For example, the difficulty in creating an association or finding relations between the items. Being able to form associations between items is thought to be an important component for retrieval. This property is thought to be how structurally coherent items are easier to process and could influence the testing effect. This will be further addressed in experiment 5.

One criticism of the studies outlined so far could be that the materials are somewhat artificial in relation to regular study materials, although the materials and results of the current experiment might be relevant to language learning processes. Therefore, a final area to explore in relation to meaningful processing in the study materials could be to see if more applied materials, similar to those studied across many topics, like expository texts, could yet elucidate the link between meaningful processing and the testing effect. In experiment five, a design similar to one used by Chan (2009) was utilised in which an interaction with the coherence of the study materials, as manipulated through the presented structure, was found with retrieval practice.

3.3 Experiment 5

In 2009, Chan found that with a retrieval-induced forgetting paradigm, structured study materials benefited more from retrieval practice than less structured study materials. In this study tested materials were compared to materials only studied once. As with findings from recent meta-analyses (Adesope et al., 2017; Rowland, 2014), the inherent structure in the information being studied can influence the magnitude of the testing effect. However, results on this topic in relation to the testing effect have been contradictory (Chan, 2009; de Jonge et al., 2015).

Previous work has shown memorial benefits for structured information over unstructured information when items are not being compared to a restudy control (M. C. An-

derson & McCulloch, 1999). Therefore Chan's findings in which increased structural coherence in the study materials benefitted more from retrieval processes is not surprising. However, more recent work that utilised a restudy control and compared the results of two different experiments found the opposite result (de Jonge et al., 2015). Experiment 5 looks to remedy the design challenges present in this previous work. For example, Chan's study did not utilise a restudy control and de Jonge et al. did not equate study time at an item level, furthermore the comparison made was across experiments. Both experiments utilised a fill-in-the-blank test for both practice and final test. Therefore, experiment 5 will be a close replication of both studies.

The aims of experiment 5 were again two-fold. Firstly, to extend the results found so far to more applied materials and secondly, to see if further work on the issue of structural coherence in the study materials would further illuminate the irregularities in the literature to date and help with answering the question to what extent meaningful processing contributes to the testing effect. In particular, experiment 5 sought to answer whether Chan's results of a high coherence test benefit would remain when a restudy control is included in the design. Or whether the results of de Jonge et al. (2015) found across two experiments, for a low coherence testing benefit would stand. The current design was marginally different from Chan's in that it did not have a within-subjects component for the coherence manipulation, instead participants were assigned to one of four conditions and was fully between subjects, which was closer to de Jonge et al. (2015)'s study.

3.3.1 Methods

Participants and Design

Participants were 95 students at the University of Plymouth, aged between 18 and 35 (M = 20.41, SD = 3.27), 86.5% female. Participants took part in the study for course credit or were paid for their time at £8 per hour, £2 each 15 minutes.

Experiment five utilised a 2 (practice task; restudy, test) x 2 (text structure; coherent, random) between-subjects design.

Sample Size Calculation

Sample size was calculated based on the effect sizes reported in (Chan, 2009), based on the main effect at delay for the unstructured data (retrieval practice versus no study control), which was f = 0.25. Calculated in G*Power (Faul et al., 2009), a sample size of 24.5 was required per group for an analysis of between-subjects factors, with two factors each with two measurements, to detect an effect of this size, with power of 0.80. This was rounded up to 25 for each group. There could have been reason to increase the sample size further given that we would expect a smaller effect size based on a restudy comparison. However, we could also expect a larger effect based on adding an extra day in delay to the design that Chan used and for utilising a between-subjects design, which could result in a larger testing effect (Rowland, 2014) therefore the sample size was kept as calculated. Due to issues in participants returning for the second session, the sample did not quite meet this requirement.

Materials

Study material. The stimulus materials for experiment five were similar to those used in (Chan, 2009). One study passage was created with information about the moon, from text on the encyclopaedia Britannica website. The final passage was approximately 900 words, which was split into four separate paragraphs of individual sentences. Paragraphs ranged between 6 and 10 sentences in length and the sentences ranged from 13 to 44 words in length. An example of a short sentence is *The Moon, is a spherical, rocky body, probably with a small metallic core, that revolves around Earth in a slightly eccentric orbit at a mean distance of about 384,000 km.* The passage was broken into these sections, as was done by Chan, in order to create the coherent and random versions of the materials. For the coherent materials, participants viewed the passage with the sentences in natural sequential order from start to finish. For the random materials, the paragraphs were presented in natural sequential order, however the sentences within each were randomised to disrupt the flow of the passage. Aside from the randomness of the structure of the sentences within each paragraph, the pas-

sages for the random passage were adapted so that the sentences could be presented out of sequential order. Each sentence was changed to make sense when presented individually. This change made the random passage twenty one words longer than the coherent passage. Full details of the study materials can be found in appendix B.2.

Test material. The test questions consisted of a 27 item fill-in-the-blank (FITB) test. The same FITB test was used for both the practice test and the final test. Test questions were made up of part of the exact sentences, or paraphrased sections of the sentences presented during the study phase. For example, for the above example sentence, the FITB item was *The moon is thought to have a small, _____ core*. Answers required were single words, in some cases a synonym for the missing word was accepted. The test questions covered each main idea that was introduced in the main passage. Full details of the test materials can be found in appendix B.3.

Procedure

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. Participants signed up to take part in both parts of the experiment. The first session took 45 minutes and the second session took around 15 minutes. Participants were tested in groups of up to six people. For the duration of each experimental session, participants were sat at a partitioned desk with their own PC. Participants wore headphones throughout session one. Besides the filler task, all elements of the experimental task were presented in PsychoPy2 (Peirce et al., 2019). Participants were told they would learn some information about the moon, that they would be tested on in a later test session scheduled for two days later. The study session consisted of two phases, a study phase and a practice phase. During the study phase each participant had twelve and a half minutes to read through one version of the passage, coherent or random. For the random version, this was a different random order for each participant. Each sentence was presented once for 25 seconds, which allowed for the sentences to be read through at least once comfortably, participants were instructed that they may have time to read through each sentence more than

once. Following the study phase, participants completed a number search (with pen and paper) for 2 minutes. A numerical filler task was chosen for the break, to minimise the likelihood of participants rehearsing something associated to the studied materials. Following the filler task, participants completed the practice phase. In the practice phase, participants in the test condition were given a FITB practice test to complete. The FITB test involved presenting part of the previously presented sentences, with one word missing, participants were instructed to fill in the blank with a word they had previously seen in the study phase, but if they could not remember the exact word, they should complete the blank with a word that conveyed the same meaning. For the final test the same FITB items were used.

3.3.2 Results

All analyses were conducted in JASP (JASP Team, 2020). All frequentist analyses where appropriate are given with the results of bayesian equivalent analyses. Descriptive statistics for the main effects of interest are given in table 3.3.

Coding Responses

As the final test required a single word response, single words were coded as either correct or incorrect. If participants had input more than one single word, but the target word was included in the phrase it was marked as correct. Spelling changes of up to two letters were coded as correct, as long as the word did not make up another meaningful word. Synonyms were included in questions that would make sense with a synonym response, for example some responses required a verb response such as braked. In which case a close synonym such as slowed, reduced or decreased was accepted as correct. Whereas, for noun target responses such a solar, only this single word was accepted as correct. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible. Please see appendix B.3, for full coding of responses.

Table 3.3
Initial and Final Test Accuracy in Experiment 5 as a Function of Practice Task and Text Structure

	Initial test		Avg	Final test		Avg
Practice task	Coherent	Random	IT	Coherent	Random	FT
Restudy	n/a	n/a	n/a	.42 (0.15)	.37 (0.18)	.39 (0.16)
Test	.50 (0.18)	.42 (0.16)	.46 (0.17)	.53 (0.16)	.43 (0.14)	.48 (0.16)

Note. The values represents mean percentages of target words recalled, SDs given in parentheses.

Main Analyses

Initial test performance. One participant did not return to take the final test, therefore their data was excluded from all analyses. This left 23 participants in the coherent test condition. Accuracy on the initial test was compared with an independent samples t-test, results did not indicate evidence for differences between the coherent test condition (M = .50, SD = 0.18) and the random test condition (M = .42, SD = 0.16), $t(45)$ = 1.77, $p > .05$, d = 0.52, although the BF_{10} = 1.02 result is inconclusive.

Final test performance. For the final test data, an initial 2 x 2 ANOVA was computed, with practice task (restudy, test) and text structure (coherent, random) as between-subjects factors. A main effect of practice task was shown, with test practice (M = .48, SD = 0.16) resulting in better performance on the final test than restudy practice (M = .39, SD = 0.16), $F(1, 91)$ = 6.62 , p = .01, $\eta_p^2 = 0.07, BF_{10}$ = 3.17.

A main effect of text structure was shown, with groups that studied the coherent text (M = .47, SD = 0.16) outperforming groups that studied the random text (M = .40, SD = 0.16) in the final test, $F(1, 91)$ = 5.67, p = .02, $\eta_p^2 = 0.06, BF_{10}$ = 2.11, although this effect was not particularly large and the bayes factor suggested the data were inconclusive.

There was no interaction found between practice task and text structure, $F(1, 91)$ = 0.39, p = .54, $\eta_p^2 < 0.01, BF_{10}$ = 0.34. Results of the final test performance are depicted

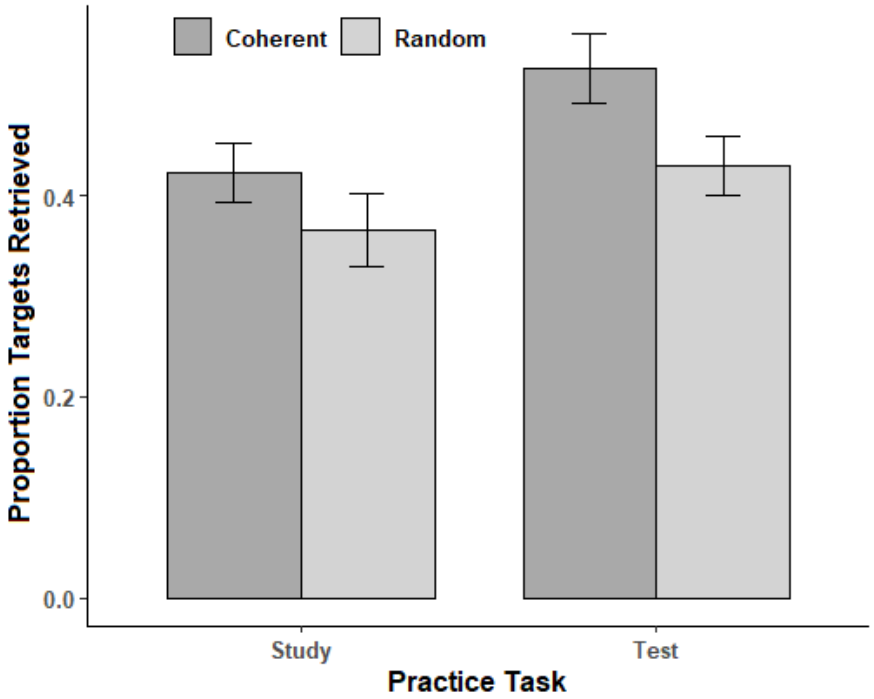

Figure 3.2. Mean target retrieval at final test as a function of practice task and text structure in experiment 5. Error bars depict the standard error of the mean.

in figure 3.2.

3.3.3 Discussion

The results of experiment five demonstrate a successful manipulation of text structure and a successful manipulation of practice task, as the coherent text was recalled to a greater extent than the random text and testing resulted in better final test retrieval than restudy practice. These results are consistent with findings that more coherent materials lead to greater retrieval and comprehension (M. C. Anderson & McCulloch, 1999; O'Reilly & McNamara, 2007) and a testing effect is found with more organised materials (Rowland, 2014).

However, as was the case in the previous experiment, there is a lack of evidence for a meaningful processing interaction with the practice task. Results therefore do not support previous findings for an interaction between practice task and text structure of the study materials that was found by Chan (2009) for a high coherence benefit or the evidence from de Jonge et al. (2015) for a low coherence benefit. Chan's study did not

offer a restudy control, suggesting that the results of this previous study might not apply to the testing effect per se. In light of the current findings it appears this is the most likely conclusion. Taking Chan's results into consideration, it might be suggested that unstructured information, without opportunity for additional restudy decays at a faster rate than structured information (Shimmerlik, 1978). However, when participants are given the opportunity to restudy the information as given in experiment 5, it seems the benefit of testing over structured and comparatively unstructured material remains the same.

However, the results of experiment 5 are not compatible with the previous results from de Jonge et al. (2015) either, whereby a low coherence advantage was seen after one week (experiment 2), where it was absent in the high coherence text (experiment 3). It could be that greater attention was afforded to the low coherent items in this study (de Jonge et al., 2015), as participants were able to self-pace their learning and the low coherent text resulted in fewer cycles through the text within the allotted time than the high coherent text. The text utilised in their study was also twice as long as in experiment 5, possibly allowing participants to further integrate the information in the high coherent text, which could have resulted in a stronger manipulation of coherence (de Jonge et al., 2015). However, as these results reflect manipulations between experiments and demonstrate evidence of different study strategies between these experiments, a manipulation within the same experiment is required with study information that reflects a stronger manipulation of coherence than the current study. Therefore, experiment 6 looked to address the reason for the discrepant results in the literature with a stronger manipulation of coherence. In experiment 6, to gain additional power, the study utilised a mixed design, whereby practice task was still manipulated between-subjects, but text coherence was manipulated within-subjects.

3.4 Experiment 6

In order to assess whether the results from experiment 5 did not just reflect a weak manipulation of coherence, a follow-up experiment was conducted which allowed for

the materials to include more cohesive materials to begin with. The main difference for experiment 6 was the use of naturally more cohesive materials, in the form of excerpts from stories that have these properties (Hakim, 2016). Arguably materials that are more cohesive would be more vulnerable to disruption to their structure, as they rely on the seamless flow from one concept to another in creating a degree of expectation and comprehension (see Hadidi and Nazerfar, 2014). Therefore, it would seem reasonable that a disruption to these highly cohesive materials might elicit results more consistent with de Jonge et al. (2015).

In addition, in an attempt to boost retrieval levels found in experiment 5, the design included the use of cue word stems during the retrieval practice task and test phase. In this way we can assess whether a disruption to more naturally cohesive materials is a stronger manipulation of the concept relating to the replication, which will further confirm the presence of the lack of interaction present in these materials.

Many effects in psychology rely on a within-subject design, for example the generation effect (Ozubko & MacLeod, 2010). Therefore, in an effort to find an interaction associated with text structure, two changes were made in experiment 5. Firstly, the meaningful processing manipulation was applied within-subjects and the practice task remained between. Secondly, the text coherence manipulation was amplified.

3.4.1 Methods

Participants and Design

Participants were 47 students and members of the public that took part in the study on campus at Plymouth University. Participants were paid for their time at £2 per fifteen minutes, or in course credit if they were registered students of psychology at the university and taking part during term time. Participants were aged between 18 and 49 years (M = 21.62, SD = 5.92), 78.7% female.

Instead of having fully between-subjects, instead a mixed design was used, whereby practice task (restudy, test) was manipulated between participants and text structure (coherent, random) of the materials was manipulated within participants.

Sample Size Calculation

Sample size for experiment 6 was calculated based on detecting a medium effect, *f* = 0.25, for a within-between ANOVA analysis, with two groups each with two measurements. Therefore a sample size of 34 in total was calculated in G*Power (Faul et al., 2009). This was increased to match the previous sample of 48 participant for the between-subject manipulation, to avoid any power issues.

Materials

The materials used were four short excerpts from four different novels or short stories. The genre of texts were chosen to be diverse from one another to maintain the interest of participants. The excerpts were found on two websites Granta and Small Beer Press. The excerpts were all capped at around 300 words long, ranging from 305 to 334 words. The passages were chosen for their cohesion on depicting a single scene or scenario within that length. Two of the four pieces included dialogue and two did not. Each passage was split into logical sentence structures of between 8 and 35 words and the presentation of each excerpt was broken into 16 sections. From those 16 sections, 16 fill-in-the-blank question items were devised. Each question related to one of the sections in the passage, however not every idea unit was quizzed, as some sections contained more than one sentence and idea unit.

For example, from the passage The Little Winter by Joy Williams, one of the sections was *Just outside Jean's town was a monastery where the monks raised dogs. Maybe she would find her dog there tomorrow.* The question item corresponding to this section was, *Gloria was to go to a monastery where monks raised d*_____. With the same question item used for both the retrieval practice task and the final test. The passages were checked for ease of reading using an online readability formula. This revealed the Flesch reading ease score to be between 74.8 and 80.3 for the four passages, one passage was easier to read and could have been understood by a young person of 8 years and the remaining three at 11 years, although all were aimed at an adult audience. This was felt to be an appropriate level to pitch the materials for a

participant sample of undergraduates to easily comprehend within a limited time frame. The amount of time allocated to read each section was matched based on its length, so that every character received the same time allocation. For the coherent presentation of the sections in the excerpt, the items were presented in their original order. For the random structure, the items were presented in a random order, which was the same during the study and practice phase, but was different for each participant. The full details of the study materials for experiment 6 can be found in appendix B.4.

Procedure

Participants were either assigned to study the excerpts or to utilise retrieval practice for each of the excerpts. Eight counterbalancing orders were established for the presentation of the four excerpts, to allow each passage to occur in one of two positions in the presentation order, and equally often as structurally coherent or random. For example, the excerpt from The Little Winter appeared equally often as either the second or fourth text and as either coherent or random presentation.

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. Participants signed up to take part in both parts of the experiment. The first session took 45 minutes and the second session took around 15 minutes. Participants attended the lab and were tested in groups of up to six people. For each experimental session, participants were sat at a partitioned desk with their own PC. Participants wore headphones throughout session one. Besides the filler task, all elements of the experimental task were presented in PsychoPy2 (Peirce et al., 2019). Participants were told that they would be learning some excerpts from four different novels and that they would complete two learning exercises for each excerpt. They were informed that the second session would be a test session and that the purpose of the session was to learn for the second session. Before the presentation of each of the excerpts, participants were asked whether they had read the novel from which the excerpt was taken. After they had answered this question they moved on to the first study phase for the excerpt.

The study stage involved the presentation of each of the sections in the excerpt one by one. The time for which each section was presented depended on the number of characters in the section, each section could be comfortably read through twice. In total each excerpt took between 3.5 and 4 minutes to present. This was slightly longer than the reading time for experiment five, in a bid to boost retrieval rates. After the initial presentation of the excerpt participants completed a number search filler task for 60 seconds with pen and paper, after which time they moved onto the learning phase for the current excerpt. A numerical filler task was chosen for the break, to minimise the likelihood of participants rehearsing something associated to the studied materials. In the learning phase, participants assigned to the restudy condition were presented the sections again but this time participants had some control over the length of time that they studied the item. Participants assigned to the retrieval practice condition completed a retrieval practice question for each of the items they had previously studied, 16 in total for each section. The retrieval practice questions were fill-in-the-blank questions, with a stem letter provided for the answer. This same retrieval practice task formed the basis of the final test phase. After participants had completed the learning phase they completed a second task for 60 seconds before moving on to the next excerpt. Participants completed all four excerpts in this fashion, before they were thanked for their time, reminded of the second session and dismissed.

For the second session participants were told that they would be answering a test for the excerpts they studied in the first session. They were told that they would need to fill-in-the-blank with a word that they had previously seen. Following completion of the final test, participants were thanked for their time and debriefed.

3.4.2 Results

All analyses were conducted in JASP (JASP Team, 2020). All frequentist analyses where appropriate are given with the results of bayesian equivalent analyses. Descriptive statistics for the main effects of interest are given in table 3.4.

Coding Responses

Items were coded blind to condition. Plurals incorrectly present or absent, obvious spelling mistakes and two letter changes to make up correct words (but not another word) were coded as correct. Due to the inclusion of cue letter stems during testing in experiment 6, very few intrusions were registered, therefore no formal analysis of intrusions was possible.

Ratings and Response Times

Participants' responses to the initial question of whether they had read the texts before were collated. None of the participants had answered yes to any of the questions, therefore it was assumed that all participants were not familiar with the texts that they studied.

Main Analyses

Two main analyses were computed, accuracy at the initial test between the two text structure conditions and the analysis of the testing effect at final test.

Initial test performance. A paired samples t-test was computed to assess the extent to which performance at the initial test differed between the two text structure types. The results of this test revealed no evidence of differences between the accuracy of the two text structure types, although the coherent text (M = .44, SD = 0.24) was retrieved to a greater extent numerically during the initial test than random text (M = .38, SD = 0.16). However, this difference was not statistically significant, $t(20) = 1.34$, $p > .05$, d = 0.29, BF_{10} = 0.50 is inconclusive.

Final test performance. A final 2 x 2 mixed ANOVA was computed with practice task as a between-subjects factor and text structure as a within-subjects factor. The results revealed no main effect of practice task, as similar performance was found for restudy (M = .35, SD = 0.12) and test practice (M = .37, SD = 0.19), $F(1, 43) = 0.26$, p = .61, $\eta_p^2 < .01, BF_{10}$ = 0.35.

Table 3.4
Initial and Final Test Accuracy in Experiment 6 as a Function of Practice Task and Text Structure

	Initial test		Avg	Final test		Avg
Practice task	Coherent	Random	IT	Coherent	Random	FT
Restudy	n/a	n/a	n/a	.37 (0.16)	.33 (0.18)	.35 (0.12)
Test	.44 (0.24)	.38 (0.16)	.41 (0.18)	.39 (0.23)	.35 (0.19)	.37 (0.19)

Note. The values represents mean percentages of target words recalled, SDs given in parentheses.

There was no main effect of the text structure, with participants performing similarly for coherent text presentation (M = .38, SD = 0.20) and random text presentation (M = .34, SD = 0.18), $F(1, 43)$ = 1.20, p = .28, $\eta_p^2 = .03, BF_{10}$ = 0.38, and no evidence was found for an interaction between these two factors, $F(1, 43)$ = 0.02, p = .90, $\eta_p^2 < .001, BF_{10}$ = 0.29. Although there were small numerical differences consistent with the pattern of results found in experiment five. Final test results are depicted in figure 3.3.

3.4.3 Discussion

The main results of this experiment echo the previous experiments in relation to the meaningful processing manipulation. There was no interaction between the two factors of interest, practice task and our meaningful processing manipulation of the structural coherence of the texts. While here no main effect for either factor is reported, the numerical differences were in the predicted pattern based on the results from experiment 5.

It is possible that despite best efforts to increase accuracy in this experiment, by presenting participants with cues, the low accuracy rates achieved are responsible for not finding a testing effect here (Rowland, 2014). However, when the results across the three experiments in this chapter are considered, alongside the variable nature of the materials employed, it seems that even where cued recall is concerned there are multiple influences to achieving a testing effect and accuracy alone is too simplistic an

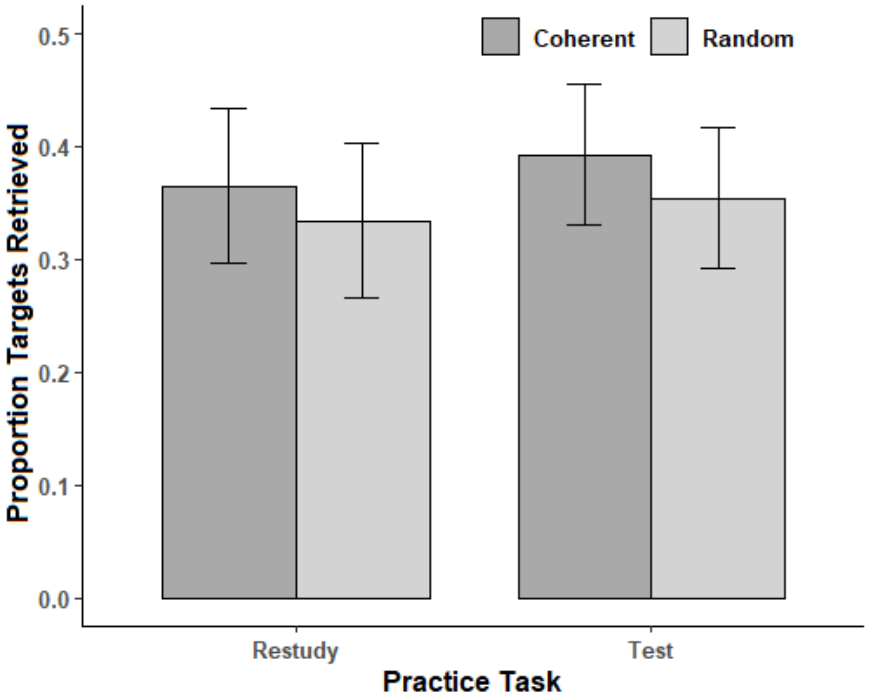

Figure 3.3. Mean target retrieval at final test as a function of practice task and text structure in experiment 6. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

explanation.

The current experiment is consistent with findings that items that are more coherent are less likely to benefit from testing (de Jonge et al., 2015; Gates, 1922; Hostetter et al., 2019). However, it is curious that the meaningful processing manipulation did not demonstrate differences in the testing effect within the same experiment here (experiments 5 & 6) and has not been demonstrated elsewhere (de Jonge et al., 2015; Gates, 1922; Hostetter et al., 2019; Roelle & Nückles, 2019; Rowland, 2014). Therefore, further work should look more precisely at the nature of coherence and cohesion and how this influences the testing effect, as once more these effects are not likely to be linearly applied (O'Reilly & McNamara, 2007).

3.5 General Discussion: Experiments 4-6

The three experiments presented here all contributed to a single aim of understanding whether the testing effect could be explained by meaningful processing of the study

materials. This was explored by the mediator effectiveness hypothesis (experiment 4) and the idea of structural coherence in text study materials (experiments 5 and 6). Findings from all three experiments here are consistent, in showing no interaction between the testing effect and the meaningful processing of the materials. While the meaningful processing manipulations in experiments 4 and 5 were successful, experiment 6 was not. However, the lack of testing effect found in experiment 6 with highly cohesive materials could be consistent with previous work that suggests highly coherent materials could be less vulnerable to interference (M. C. Anderson & McCulloch, 1999), less memorable (Hostetter et al., 2019) and possibly less likely to benefit from testing (de Jonge et al., 2015).

In experiment 4, a more direct examination of the mediator effectiveness hypothesis was conducted. The materials were manipulated to enable a greater opportunity to establish links between the cue and target word in the helpful mediator trials (English definition), compared with less opportunity in the less helpful mediator trials (Swahili definition). While a main effect of the meaningful processing manipulation was seen on final retrieval rates, this did not interact with the testing effect. In fact, there was evidence for no testing effect and even a negative testing effect for the no feedback test condition, despite the fact that the meaningful processing manipulation was present and strong. Although, it is not especially common to find no testing effect, some literature does report that when effort is required to integrate the materials (Chen, Castro-Alonso, Paas, & Sweller, 2018; Van Gog & Sweller, 2015), retrieval rates are low (Rowland, 2014) or when testing is immediate (Roediger & Karpicke, 2006b), testing is less beneficial. The results of experiment 4 are consistent with this. However, as experiment 4 achieved similar retrieval rates to experiment 5 where a reliable testing effect was found, low accuracy alone does not explain the findings. Support for the elaborate retrieval hypothesis could be taken from the results of experiment 4, based on the fact that elaborate restudy performed at similar levels to the test with feedback condition. However, as this condition, also outperformed the test only condition and performed at similar levels to the restudy only condition, this makes for a less straightforward inter-

pretation. Previous work has struggled to find support for the ERH when an elaborate task has been compared to retrieval practice (Lehman & Karpicke, 2016; M. A. Smith, Blunt, Whiffen, & Karpicke, 2016). In contrast to those results, the current results suggest that at least sometimes elaboration is as beneficial or more beneficial than testing. Consistent with previous work (Willoughby et al., 1994), this could yet depend on the nature of the study materials, as McDaniel et al. (2009) found that an elaborate study task (note-taking) performed similarly to retrieval practice (read-recite-review) when the study materials were more complex.

The results of experiment 4 do not provide convincing evidence in favour of the mediator effectiveness hypothesis, as the test groups failed to outperform the restudy groups in the helpful mediator condition. No increased benefit was seen either based on the test only condition or the test with feedback condition for the helpful mediator (English definition) condition over either restudy or elaborate restudy. Although the test with feedback condition did not show improved performance over the restudy condition for the helpful mediator trials, there was a numerical trend in this direction. These results suggest that mediating information alone is not useful to the direct effects of testing, possibly due to an absence of the mediating effects associated with a repeat restudy-retrieval design, on retrieval practice. For example, the benefit of retrieval practice has previously been enhanced via improved encoding following retrieval practice (Kornell et al., 2009). As the investigation for this thesis is concerned with the impact of meaningful processing on the direct effects of retrieval practice, it was beyond the scope of the current thesis to further investigate the mediator effectiveness hypothesis via multiple retrieval practice and restudy opportunities. It will be for future research to assess whether the benefit associated with combining mediators with retrieval practice, was previously only demonstrated due to the high levels of repeated practice (Pyc & Rawson, 2010) and the boost to testing effects associated with this (Rowland, 2014).

In experiment 5, the structure of the materials was manipulated between subjects, at final test again a main effect of practice task was found as well as a main effect of meaningful processing, but no interaction was found. This study was designed to

address the issues of previous work in this area (Chan, 2009; de Jonge et al., 2015). The results strongly suggest that the inclusion of a restudy control task in experiment 5 as opposed to no study control task in Chan's study contributed to finding no interaction between the structural coherence of the study materials and the testing effect. The results of experiment 5 once more indicate that changes in the meaningful processing of the study materials do not impact the magnitude of the testing effect.

Experiment 6 went some way to address the issues present in experiment 5. For example, an attempt was made to boost retrieval practice by providing additional cue stems during the retrieval practice task and the final test. Whilst the this did not boost overall retrieval, again there was no interaction between practice task and text structure. There were no main effects present for either factor, however the numerical differences were in the predicted pattern based on the results from experiment 5. Suggesting that a lack of interaction, merely demonstrates that the interaction is likely to be absent with manipulations of this nature. In addition, it could also be that the testing effect is not likely to be present with highly cohesive materials such as novel extracts (Hostetter et al., 2019).

One point to note about the results from experiments 5 and 6 is that the materials were artificially manipulated to alter text coherence and this did not impact the testing effect. There were also likely differences in text cohesion, or how reliant concepts are on one another for their comprehension between the two experiments. It is possible that the inter-relatedness of the concepts altered whether testing was beneficial, as evidence by a main effect of practice task seen in experiment 5 but not 6. However, the results of the meta-analysis by Rowland (2014) did not suggest this to influence the testing effect, when study materials reflected differences in cue-target relations, rather than as a feature of text materials. While text cohesion across experiments 5 and 6 did seem to alter whether a testing effect was found and has recently been found to influence when testing is beneficial (Roelle & Nückles, 2019), more work in this regard is needed as text cohesion manipulations are yet to be explored in the same experiment.

Evidence from previous work (Rowland, 2014) has shown that aspects of organisation do influence the testing effect, but that this is typically with larger differences, individual words compared to prose passages. The nature of the studies included in Rowland's analysis likely represent materials of the nature seen in experiment 5, not 6, which could be where the findings are more reliable. While the materials used in experiment 6 might appear to be less applicable due to the fact that most educational topics handle materials less cohesive than those used in experiment 6. Topics in the humanities can often require story-like or narrative information to be studied verbatim. It might therefore be an opportunity for future work to exploit the lack of testing effect seen with these materials, to further understand when testing will be useful. It is with a degree of caution that this is suggested as the influence of cohesiveness in texts can be a complex affair (O'Reilly & McNamara, 2007), and its relationship with the testing effect might not be an exception to this.

To their merit, the three experiments described here, utilised different designs, from mixed designs in experiments 4 and 6, to fully between-subjects in experiment 5. As previously documented, design aspects (Mulligan et al., 2016; Rowland, 2014; Rowland, Littrell-Baez, Sensenig, & DeLosh, 2014) could inform the magnitude of the effect. As we find a consistent lack of evidence for the influence of a meaningful processing manipulation in aspects of the materials across these designs, the evidence clearly suggests an absence of this effect. At the very least, the results of chapters 2 and 3 show that differences in meaningful processing of the study materials do not make a key contribution to the testing effect.

However, it must also be noted that there are differences in the structure of the materials across the three experiments in chapter 3, which in two cases have prevented a testing effect being observed. While it is not immediately clear what aspect of the materials might be influencing this, as previously noted it is possible that ease of integration could be responsible (Van Gog & Sweller, 2015), perhaps based on processing or reading fluency (de Jonge et al., 2015), which might be useful to measure on an individual level in explaining the results (O'Reilly & McNamara, 2007). In line

with meta-analytic results (Adesope et al., 2017; Rowland, 2014), any differences in the magnitude of the testing effect associated with the structure of the study materials likely depends on a more dramatic manipulation of meaningful processing. This can be seen as a positive result for the application of these results, as small changes in meaningful processing in the study materials do not alter how helpful testing will be. Meanwhile, there is still difficulty generally in retrieving less meaningful materials and future work might look to show how memory for these items might be boosted. In line with this, previous work has demonstrated that individuals tend to give most processing resources to items that are in the mid range of difficulty from a selection (Metcalfe & Kornell, 2003).

To further add to the evidence given in chapters two and three in relation to meaningful processing in the study materials contributing to the testing effect, a mini meta-analysis was conducted. Previous work has suggested that mini meta-analyses are useful for further interpretation of results and increasing the precision of findings, particularly where results might be under-powered or include null findings (Goh, Hall, & Rosenthal, 2016). A mini meta-analysis was conducted in JASP (JASP Team, 2020) using the meta-analytic function. The analysis was a Hedges random effects model computed across all six studies from chapters two and three to determine the effect of meaningful processing on the magnitude of the testing effect, weighted by the size of the sample in each experiment. The mini meta-analysis assessed the aggregated testing effect in the presence of two moderators; delay period (immediate versus delay) and meaningful processing (high meaningful processing versus low meaningful processing). Based on the fact that previous research has sometimes made opposing predictions on when meaningful processing will benefit testing, higher accuracy conditions were labelled as high in meaningful processing and lower accuracy conditions were labelled as low in meaningful processing. The heterogeneity indicated in the sample was I^2 = 42.78, indicating a below medium (< 50%) level of heterogeneity between samples (Higgins & Thompson, 2002). This statistic indicates that 43% of heterogeneity between the effects in the analysis is not accounted for by sampling error

within each effect (Borenstein, Higgins, Hedges, & Rothstein, 2017). The main estimated effect of practice task (test advantage over restudy) was not significant, p=.68, d = 0.07 [-0.26,0.41]. The heterogeneity between samples was further explored in the moderator analyses. A significant amount of variance could be explained by the delay moderator, p <.02, d = 0.51 [0.10, 0.91], showing no testing effect for studies with an immediate delay d = 0.10 [-0.35, 0.56], but a significant testing effect for studies with a delay greater than one day, d = 0.52 [0.14, 0.91]. There was no effect of meaningful processing p = .33, d = -0.18 [-0.58, 0.22], showing insignificant effects of testing for both low meaningful items d = -0.16 [-0.58, 0.26], and high meaningful items d = 0.12 [-0.26, 0.50].

The results demonstrate that across this sample, the delay to the final test rather than how meaningful the processing of the items is, indicates the magnitude of the testing effect. These results confirm that meaningful processing of the study materials alone is not a contributor to the testing effect. The results show that alternative approaches, such as examining how cohesive or easy to integrate text material is, or when feedback is most useful in relation to specific materials learned are likely to be useful approaches to future work looking at how differences in encoding of materials contributes to the testing effect.

Chapter 4

Meaningful processing during retrieval practice

Chapter four diverges from the focus of the previous chapters, moving away from meaningful processing in the study materials to meaningful processing during the retrieval practice task and learning outcomes. In assessing how differences in processing during the retrieval practice task influences the magnitude of the testing effect, the focus is more applied and aims to further understanding of when retrieval practice is most beneficial. In addressing the issue of meaningful processing in the learning outcomes, the current chapter applies the controlled design of the previous studies to an area of the testing effect literature that has been under-explored with this level of concern, transfer learning.

4.1 Introduction

The previous two chapters examined the role of meaningful processing in relation to the testing effect based on relevant work in the literature. Namely in relation to the elaborate retrieval hypothesis, the mediator effectiveness hypothesis and in relation to the structural coherence of the study materials. Across six experiments there was no evidence to support the role of meaningful processing in the testing effect based on the properties of the materials being studied during encoding. However, as outlined in the introductory chapter, meaningful processing can take many forms and therefore it is useful to explore this concept in different ways to understand its impact on the testing effect and where future work would be best focused in this regard.

Some work has shown that meaningful processing during the practice task is a useful focus in relation to the testing effect, with the potential to alter the effectiveness of retrieval practice (Endres et al., 2017; Larsen, Butler, & Roediger, 2009). However,

direct work in this area (manipulating differences in the practice task as the focus) has concentrated on transfer learning (Hinze et al., 2013), which is a slight departure from the testing effect format utilised in this thesis so far. Transfer learning in relation to the testing effect, occurs when the retrieval practice test and final test are assessed differently. Transfer therefore is a broad term, which encompasses all from changes in test type between initial and final test, to changes in the test items and applying knowledge tested to new solutions (Pan & Rickard, 2018). As meaningful learning can be thought of in relation to whether knowledge transfer occurs (Mayer, 2008), chapter 4 will assess two areas of meaningful processing, in relation to processing that occurs during the retrieval practice task and the learning outcomes of this processing.

Experiment 4 in the previous chapter highlights that different forms of processing during the practice task influence memory performance. In experiment 4 the elaborate study practice task showed similar accuracy at final test to the retrieval practice task with feedback and superior performance than the test only condition. In line with the results from experiment 4, previous research has shown that more meaningful processing during a practice task contributes to performance.

For example, Coane (2013) found that a deep processing task, similar to the elaborate restudy practice task used in experiment 4, was more beneficial than restudy for memory after a delay of 10 minutes and 2 days for both younger and older adults. Unlike the results of experiment 4, the elaborate processing task was not as beneficial as the retrieval practice task, which included feedback however. Other studies have found similar results (Blunt & Karpicke, 2014; Karpicke & Blunt, 2011; Lechuga et al., 2015; McDaniel et al., 2009), although McDaniel et al. (2009) found that the elaborate study task (note-taking) performed at similar levels to the retrieval practice condition (read-recite-review) when the study materials were longer and more complex in experiment 2 than in experiment 1. Karpicke and Blunt (2011) compared retrieval practice to an elaborate study task known as concept mapping and found retrieval practice resulted in superior final test performance one week later. Yet once again this more meaningful form of practice task outperformed a study only control. These results show that ad-

ditional processing during a study practice task is beneficial for retention, posing the question of whether additional meaningful processing during the retrieval practice task also serves to boost retention and in turn the magnitude of the testing effect.

Relevant to this question is the practical issue of which retrieval method is most efficacious in the testing effect. Studies have compared the standard retrieval practice tasks and show mixed results for any differences. For example, while some studies have found that MCQs during practice are more beneficial than short answer questions (Greving & Richter, 2018), other studies have found the opposite trend (Butler, Karpicke, & Roediger, 2007; Larsen et al., 2009), or equivalent performance between them (M. A. Smith & Karpicke, 2014). However, there is also a debate as to whether there are differences in the processing required by these different retrieval practice strategies. With questions over whether free recall for example, offers a more difficult processing task than cued recall or whether MCQs are easier to process than a cued recall task. Yet others have suggested that MCQs could be more difficult than cued recall or short answer questions, because not only do participants need to recall the correct answer but they also need to remember why the other, often viable, answers are not correct. The results of these studies and surrounding debate pose yet unanswered questions of the desirable difficulties account and therefore others, such as the elaborate retrieval hypothesis (Carpenter, 2009) and the episodic context account (Karpicke et al., 2014), where difficulty forms a tenet of the theory.

Perhaps rather unexpectedly in light of these results, Rowland (2014) found that cued recall tests tend to benefit the testing effect to a similar (high exposure sample) or greater extent than free recall (total sample) and MCQs. So while some evidence suggests that different types of retrieval tasks do alter retention and the testing effect, very few studies have directly examined how differences in meaningful processing during retrieval practice impacts the testing effect. Those studies that have assessed this, have been motivated by greatly different perspectives to do so.

One such study explored this from the perspective of the elaborate retrieval hypothesis of the testing effect by adding an elaborate component to the retrieval practice

task. Endres et al. (2017) found that an elaborate retrieval practice task that encouraged participants to relate the retrieval practice material to their own experiences, was comparable to a free recall retrieval practice task on a short answer test one week later. In addition, authors looked at two forms of retention, detail and comprehension scores and found that comprehension scores were predicted by the amount of elaboration that occurred during retrieval, regardless of the practice task. This suggests that more meaningful processing during retrieval can be instrumental to effective learning outcomes.

Similarly, Larsen et al. (2013) assessed different forms of practice from an applied perspective. Results also showed that practice was beneficial when it was personal. In this study different methods for learning medical diagnosis protocols were assessed and clinical experience was the most beneficial overall practice task when learning was assessed after 9 months. Furthermore, M. A. Smith et al. (2016) found that transfer learning could be achieved with free recall or prompted recall, suggesting that processing that helps to organise knowledge in a meaningful way (Hinze et al., 2013) is useful to long-term memory and possibly the testing effect. However, in each of these examples the outcome measures are forms of transfer measures, as the final test contains different items and test types to the practice test. Therefore no studies have stringently assessed the effects of more meaningful retrieval on specific retention when the initial and final test conditions are matched. This is an important point, because the specific effects of retrieval practice are larger (Rowland, 2014) than the transfer effects of retrieval practice (Pan & Rickard, 2018). [1] Therefore more meaningful processing might be uniquely helpful to specific effects of retrieval practice, which to date has not been explored.

The explanation of the benefits of a more meaningful retrieval task have been suggested as a result of research focused on transfer learning. Hinze et al. (2013) demonstrated that the focus adopted during retrieval practice is an important component to

[1] The current estimate of the specific effects of retrieval practice, as assessed through Rowland (2014)'s meta-analysis is thought to be g = 0.50, 95% CI [0.42, 0.58], whereas the suggested benefit of transfer learning through retrieval practice is thought to be somewhat smaller, d = 0.40, 95% CI [0.31, 0.50].

what information will be retained. Participants were assigned to a particular condition and received an example text, then worked on a target text and subsequently received instructions on the retrieval practice task in line with this. For example, participants were provided with an example text with questions that required a detail focus or an inference focus. For the target text participants were told to retrieve the information previously studied to be able to answer the types of questions they had previously answered on the example text. On the final test, participants were tested on both detail and inference questions. Participants who had previously answered inference questions for the example text had greater inference retrieval on the final MCQs than participants who completed a non-specific free recall task and those who were expecting detail questions. They also found that the focused retrieval group led to better quality responses in their retrieval practice task. The explanation for the findings is that retrieval practice tasks that require the individual to reconstruct their knowledge to a greater extent appear to benefit more from testing.

Other studies specifically assessing transfer have shown that meaningful forms of retrieval can yield large transfer effects. For example, Butler (2010) found that repeated retrieval practice with elaborate short answer questions, promoted beneficial transfer to new questions when compared to a restudy control. However, it is worth noting that this study involved repeated cycles of retrieval practice with accompanying feedback, which we know inflates the utility of retrieval practice (Exp 4 results, chapter three; Rowland, 2014). Indeed, van Eersel, Verkoeijen, Povilenaite, and Rikers (2016) found that with a replication based on Butler's application transfer experiment (experiment 3), that feedback significantly contributed to the ability for test practice to show benefits for application transfer. This is another case that highlights that a less constrained approach has been taken in relation to assessing the benefit of meaningful forms of retrieval practice.

One recent study found that how integrated the items were during the learning phase and subsequent retrieval practice phase affected how much information was retained (Eglington & Kang, 2018). However, as the results of chapters 2 and 3 demon-

strate, differences in meaningful processing of the materials during the learning phase are not likely to be responsible for these results. Therefore, once more there is suggestion that meaningful retrieval practice should be beneficial to the testing effect.

As already highlighted, the research in this area has been approached in a less systematic way. In this way, possibly as motivation for these studies typically lies in the applied domain, a common feature of the literature featured here is that adequate control study tasks are often missing. In addition, different forms of practice task have typically been compared, which can lead to different amounts of initial retrieval accuracy. Research carefully controlling the nature of the retrieval practice task and comparative study control tasks has not been undertaken and is necessary to further understand the utility of meaningful processing in the retrieval practice task.

Results of the literature reviewed above highlight that: 1) Few studies have explored differences in processing during the retrieval practice task. 2) Many studies that have looked at meaningful processing in the retrieval practice task are seen in relation to transfer results. 3) Retention testing effects are typically larger than transfer testing effects. 4) Studies that compare different retrieval practice tasks, often do not attempt to match the restudy control task. 5) Results have seldom addressed the direct effects of transfer, without feedback and repeated test cycles.

In addressing these points this chapter will explore whether meaningful processing during the practice task influences the testing effect for specific information (experiment 7 and 10). In addition, I will also explore the result of this meaningful processing on meaningful learning outcomes (experiment 8 and 9), in relation to transfer learning (Barnett & Ceci, 2002), where meaningful processing during the retrieval practice task has shown the most promise to date. The experiments in the current chapter will build on aspects explored in experiments 5 and 6, by continuing to utilise more educationally realistic study materials. Cued recall will once more be used for its known ability to produce a testing effect, for its comparison to the previous experiments and to be able to restrict the variation in target answers. Experiment 7 was designed to address the issue of whether differences in the amount of meaningful processing achieved during the

retrieval practice task impacts the magnitude of the testing effect for specific retention.

4.2 Experiment 7

Experiment 7 aimed to address how differences in meaningful processing during retrieval practice contribute to the testing effect. Something that has not been explored and could be relevant to the specific effects of testing is whether providing an opportunity to elaborate on information relating to the target increases memory for the target. For example, in comparing elaborate retrieval (self-relating the study materials) to non-elaborate retrieval, Endres et al. (2017) concluded that although the condition with additional elaboration during retrieval performed at comparable levels to test, the time taken suggested that it might not be as useful a strategy. Yet, if the additional construction of answers, also show a benefit to transfer learning then there would be an added benefit beyond specific retrieval. This is important as it might increase our understanding of when retrieval needs to be specific and when an additional benefit of elaboration need not impede the recall of specific targets.

Earlier work by Soraci et al. (1994), demonstrated that generating information in response to a cue was helpful for later item recall over restudy. This has subsequently become know as the generation effect, which relies on semantic memory rather than episodic memory, but shows that generating information is more retrievable at a later time-point than restudying it (Bertsch, Pesta, Wiscott, & McDaniel, 2007). Furthermore, this idea has been supported more recently in relation to the transfer testing effect. In a meta-analysis on transfer effects, Pan and Rickard (2018) found that elaborated retrieval practice, whereby the retrieval practice incorporates broad encoding, for example broad retrieval instructions and explanatory recall, influences the likelihood that transfer testing effects will be found. However, authors recognised that only a small number of broadly categorised studies contributed to the result and suggest more work is required in this area. With work showing that generation effects and testing effects share some common phenomena (Mulligan & Peterson, 2015), it is possible that factors associated with generation are also at work in the testing effect literature.

Based on this evidence we could suggest that in line with an elaborate retrieval approach, when the retrieval practice task allows for additional information to be generated at the time of retrieval, this additional information could serve as later cues to final test retrieval. Importantly, experiment 7 served to provide matched restudy conditions, where previous studies have failed to do so (Endres et al., 2017).

Previous empirical research leads to a prediction that more meaningful processing at the time of retrieval practice will result in a larger testing effect than the less meaningful processing at the time of retrieval practice. This prediction is compatible with the elaborate retrieval hypothesis, the desirable difficulties account and the episodic context account.

4.2.1 Methods

Participants and Design

Participants were 56 students and members of the public that took part in the study on campus at Plymouth University. Participants were paid for their time at £2 per fifteen minutes, or in course credit if they were registered students of psychology at the university and taking part during term time. Participants were aged between 18 and 35 years (M = 23.2 , SD = 5.3), 78% female. The present study utilised a 2 (practice task; restudy, test) x 2 (practice type; *what practice*, *what & why practice*) mixed design, with practice task as a between-subjects factor and practice type as a within-subjects factor.

Sample Size Calculation

The calculation for the sample required for this particular experiment was conducted in G*Power. As there was no reliable marker for the size of the effect to expect. The sample size was designed to detect a medium-sized main effect (Rowland, 2014), as cued recall effect sizes tend to be larger than this g = .61. The G*Power (Faul et al., 2009) analysis was based on an ANOVA analysis with two between group factors and two within group factors. This gave each group size sample of 24.5 participants, with power of 0.80. Each group was rounded up to 28 participants. As the current study

required 2 within-subjects groups, the final sample was 56 participants.

Materials

Materials were facts relating to different animals; sharks, penguins and crocodiles. The information was taken from various websites. Study statements were constructed that included both a plain fact (*what* element) and an explanation for that fact (*why* element). Twelve study statements made up the materials for each animal. For example, one of the penguin study items was *Today, wild penguins exhibit no particular fear of human tourists (what element), this is because they are not used to danger from animals on solid ground (why element).* In addition, twelve questions were constructed for the *what* element that asked for the answer to be a direct object noun present in the fact and twelve follow-up questions were constructed for the *why* element, which mostly consisted of the words "why is that?" These two types of questions made up the *what* practice questions and the *why* practice questions respectively. Equivalent items made up the restudy practice items, which included only the *what* element from the original study item or both the *what & why* element presented together which was the same as the full original study item.

In the practice phase half of the items were practiced as the whole item as previously presented in the study phase, which included both the *what & why* element either as restudy or test. The other half of the items were practiced with only the *what* element of the item previously presented in the study phase. For each set of twelve facts the order of practice trial was counterbalanced, so that odd trials were practiced as one type and even trials were practiced as the other. The trial order was counterbalanced across participants. The full details of the study materials for experiment 7 can be found in appendix C.1.

Procedure

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. Participants signed up to take part in both parts of the experiment. The first session took 30 minutes and the second session took

around 15 minutes. Participants attended the lab and were tested in groups of up to six people. For each experimental session, participants were sat at a partitioned desk with their own PC. Participants wore headphones throughout session one. Besides the filler task, all elements of the experimental task were presented in PsychoPy2 (Peirce et al., 2019).

Participants were instructed that they would learn some information about animals in the first session and be tested on what they had learned during the second session. Participants were randomly allocated to condition, which was either restudy practice or test practice for all three animals.

All participants learned about three different animals in succession in the same order in session one. For each animal there was two learning phases, a study phase and a practice phase. Prior to the study phase, participants were told which animal they were studying and asked to rate how familiar they were with the animal and how interested they were in the animal on a five point scale. For each animal, participants were first presented with the twelve study statements, each of which included the what and why element of the item together. Each statement was presented for 15 seconds. This gave participants ample time to read the statement multiple times. Following the study phase, participants were again asked to rate both their familiarity and interest in the information they had studied. Following the post-study rating for each animal, participants were given a 60 second filler task in which they completed a puzzle. Participants had the option to complete a number search or sudoku puzzle during the break (both puzzles supplied on double-sided sheet of paper). A numerical filler task was chosen for the break, to minimise the likelihood of participants rehearsing something associated to the studied materials. Following the filler task, participants entered the practice phase, which was different based on the between-subjects manipulation, restudy or test practice.

For restudy participants, the presentation of the statements they had already seen was altered for half of the trials. For half of the trials for each animal, only the what element was presented, the remaining trials presented the full statement as was seen

in the study phase (what and why element together). Participants were given 10 seconds to restudy the what elements and 20 seconds to restudy the full statement seen previously. The same amount of practice time was provided for the equivalent manipulations in the test condition during the practice phase. The order of these statements was counterbalanced by item and participant. Participants were told to pay special attention to the words used as they would be asked to give exact words from the sentence in a future test. For the test condition participants, there was a comparable alteration during the second learning phase. Test participants received short answer test questions during the practice phase, participants were asked a what question for half of the study items and a what & why question for the other half.

For the what questions a direct noun object was required for the answer. For the why part of the question, participants were required to recall as much of the why element of the study item previously paired with the what element. Participants therefore saw all of the original plain facts with a one word direct object noun missing that they were asked to recall, but in addition, for half of the items they were also asked to retrieve the "why" explanation associated with it and presented alongside the item in the study phase. For what elements, participants were instructed to try to retrieve the exact word that was missing from the sentence and that it would only be one word that was required in the answer for these cases. For the "what" questions, participants were given 10 seconds to answer. For the what questions combined with a "why" follow-up component, participants were given 20 seconds to answer. Following the second learning phase, each participant completed another filler task for 60 seconds, before moving on to the next animal. Participants completed the same practice task across all three animals. On the final animal, there was no final filler task after the practice phase. Participants were thanked for their time and reminded of the time and date of the second session that was scheduled.

Between three and six days later (M = 3.45), all participants returned to complete the final test. The final test consisted of 36 questions. All of the questions were for the "what" element previously quizzed in the same way for test practice participants. The

final test was self-paced. Following completion of the test session, participants were debrief and thanked for their time.

4.2.2 Results

All analyses were computed in JASP (JASP Team, 2020) for both frequentist and equivalent bayesian tests (given where appropriate). Descriptive statistics for the main effects of interest are given in table 4.1.

Coding Responses

The responses of the retrieval practice group during the practice phase were coded based on the accuracy of the retrieved missing noun object. Participants' answers could be coded as correct based on the exact answer, exact answer with spelling correction or if correct answer was provided in a phrase or close synonym of the correct answer provided. Synonyms were taken from merriam webster online or the agreement of the coders. The same scoring criteria were used for responses at the final test. An agreement on the criteria based on dictionary synonyms and opinion was agreed upon for a subset of five participants initially coded. This same criteria was applied to the remaining answers. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible.

Ratings and Response Times

Participants ratings prior to study and after study were averaged across the three animals to make one pre and post study rating for interest and one pre and post study rating for familiarity for each participant. Due to a logging error, the ratings for two participants in the restudy condition were not collected and so are not included in the comparison tests given here. Two independent-samples t-tests were computed for ratings given both prior to and following study practice, for the familiarity and interest score. This was to establish that there was no difference in familiarity or interest in the information being studied either prior to study or following study practice between the restudy and test conditions. Results showed that prior to study, both the restudy (*M*

= 2.70, *SD* = 0.98) and retrieval practice condition (*M* = 2.30, *SD* = 0.89) showed no evidence for differences in familiarity with the materials, $t(52)=1.56$, $p = .13$, $d = 0.42$, $BF_{10} = 0.74$. For interest ratings the pattern was the same, with restudy (*M* = 3.85, *SD* = 0.73) and retrieval practice groups (*M* = 3.70, *SD* = 0.64) , $t(52)=0.76$, $p = .44$, $d = 0.21$, $BF_{10} = 0.35$, showing similar levels of interest, with just outside moderate evidence for the null.

In addition, further tests on ratings were conducted to assess how familiarity and interest ratings compared before and after the study phase in order to ensure that participants were not overly familiar with the facts and did not lose interest throughout the experiment. A Wilcoxon signed-rank test (due to violation of normality) revealed that participants rated their familiarity prior to learning the study materials as higher (*Mdn*=2.33) than after studying the materials (*Mdn*=1.67), $T=945.50$, $p<.001$, $r=.27$. Suggesting that the facts were not familiar to them. The paired samples t-test for the interest ratings revealed participants rated their interest prior to studying the materials as lower (*M*=3.77, *SD*=0.68), than after studying the materials (*M*=4.19, *SD*=0.57), $t(54)=-5.53$, $p<.001$, $d=-0.75$, $BF_{10} = 16313$. Suggesting that the information held their attention.

Main Analyses

Initial test performance. A paired samples t-test compared the initial test accuracy between the explanatory trials (*M* = .69, *SD* = 0.15) and the fact trials (*M* = .66, *SD* = 0.18) for the “what” items, results showed no evidence of differences between the groups levels of accuracy during the retrieval practice test, $t(27)=1.23$, $p = .23$, $d = 0.23$, $BF_{10} = 0.40$. During the retrieval practice task 86% of "why" answers contained a response of some form and 14% were left blank. Answers that were left blank contained both items that were registered as not recalled (79%) or having timed out before a response could be registered (21%).

Final test performance. A 2 x 2 mixed ANOVA was computed for the number of items recalled accurately at final test. This was accuracy for the “what” item only,

Table 4.1
Initial and Final Test Accuracy in Experiment 7 as a Function of Practice Task and Practice Type

	Initial test		Avg	Final test		Avg
Practice task	What	W & why	IT	What	W & why	FT
Restudy	n/a	n/a	n/a	.57 (0.17)	.52 (0.20)	.55 (0.17)
Test	.66 (0.18)	.69 (0.15)	.67 (0.16)	.65 (0.19)	.69 (0.16)	.67 (0.17)
Avg FT	n/a	n/a	n/a	.61 (.20)	.61 (.18)	n/a

Note. The values represents mean percentages of target words recalled, SDs given in parentheses.

with practice task (restudy, test) as a between-subjects factor and practice type (*what*, *what & why*) as a within-subjects factor. Results revealed a main effect of practice task, with test practice accuracy (M = .67, SD = 0.17) being greater than restudy practice accuracy (M = .55, SD = 0.17), $F(1,54)$ = 7.26 , $p < .01$, $\eta_p^2 = .12, BF_{10}$ = 5.64.

There was no main effect of practice type, with *what & why* practice trials (M = .61, SD = 0.20) overall showing similar accuracy to *what* practice trials (M = .61, SD = 0.18), $F(1,54)$ = 0.01, $p < .01$, $\eta_p^2 < .01, BF_{10}$ = 0.20. There was some evidence for an interaction between these two factors, $F(1,54)$ = 5.82, $p < .01$, $\eta_p^2 = .10, BF_{10}$ = 2.44.

Follow-up one way ANOVAs revealed no evidence for differences in the final accuracy for the *what* practice trial format between restudy (M = .57, SD = 0.17) and test practice (M = .65, SD = 0.19), $F(1,54)$ = 2.69, $p < .11$, $\eta_p^2 < .05, BF_{10}$ = 0.82. However, there was a difference found in final accuracy for the *what & why* practice trial format between restudy (M = .52, SD = 0.20) and test practice (M = .69, SD = 0.16), $F(1,54)$ = 11.55, p = .001, $\eta_p^2 = .18, BF_{10}$ = 25.29. Results of the final test are depicted in figure 4.1.

4.2.3 Discussion

In experiment 7 it was shown that altering the format of the practice task can influence the magnitude of the testing effect. Results showed that when the retrieval practice task required only recall of the *what* element, limited advantage was seen at final test

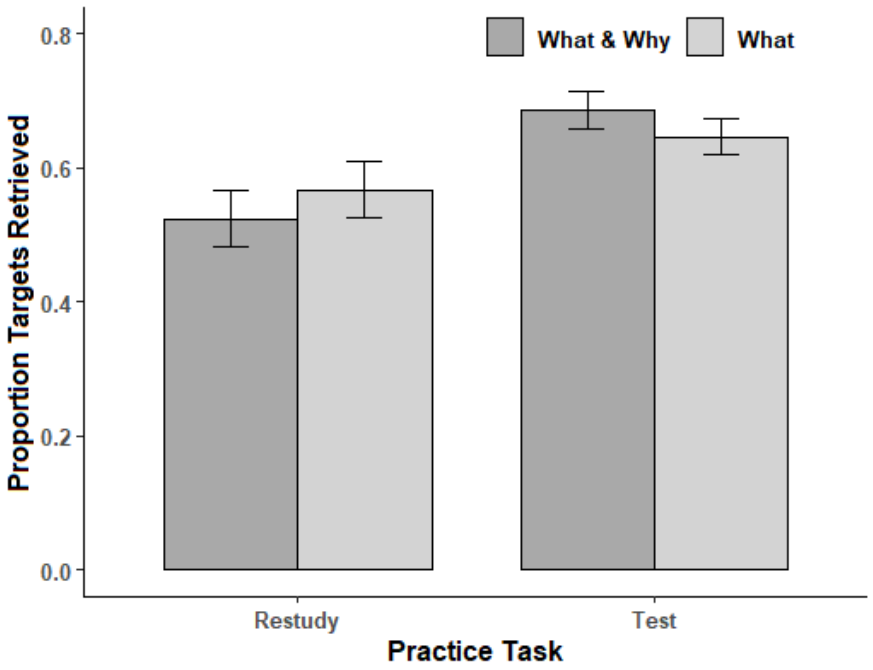

Figure 4.1. Mean target retrieval at final test as a function of practice task and practice type in experiment 7. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

compared to an equivalent restudy task. However, when the practice task required more elaborate recall of the *what & why* element, retrieval practice was more advantageous to final test accuracy than a restudy equivalent. There are a number of points to discuss in relation to this.

It is possible the difference found between the restudy and retrieval practice conditions for *what & why* trials could have been due to the additional explanatory information, the *why* element, being advantageous in the testing condition but not so in the restudy condition. This could be because the retrieval practice task enabled additional retrieval cues to be generated (Pan & Rickard, 2017), while the matched restudy task did not. Interpreting how this evidence supports existing accounts is challenging, as it can support all of the main accounts. For example in support of the elaborate retrieval hypothesis (Carpenter, 2009), during retrieval practice by actively retrieving more information, in the *what & why* practice trials, additional cues could have been generated, more elaborate cues that better guided retrieval at final test.

But these ideas equally apply to the episodic context account (Karpicke et al., 2014). As the *why* element retrieval in the *what & why* trials allowed for more context reinstatement from the original study episode than the *what* only trials, possibly enabling more updating to the retrieval context and thereby providing additional retrieval cues at final test. These results are also consistent with a desirable difficulties (E. L. Bjork et al., 2011), based on the fact that the *what & why* trials involved more effort than the *what* only trials, with both the time taken on the task and the amount of effort required to complete the task. With a matched restudy control task it would appear that increased effort during the practice task increases the magnitude of the testing effect. Furthermore, these results also fit with the explanations from the transfer literature, that constructing knowledge (Hinze et al., 2013) is beneficial to the testing effect in a broad sense. The opportunity to construct knowledge from the previous study session was greatest in the *what & why* trials and was more beneficial than an equivalent restudy trial than less opportunity to construct knowledge.

While the results are consistent with current theory and phenomena of the testing effect, being able to dissociate the effects associated with each of these concepts would be a challenge. Although the comparison between the two test conditions and the two restudy conditions show the differences to be small, it is possible that with more diverse retrieval practice tasks and possibly study materials the true nature of these differences could be assessed.

Interestingly, recent work (Roelle & Nückles, 2019) found that generative, more elaborate learning that involved linking concepts to existing knowledge was most beneficial when the information being learned was less cohesive. It seems that linking new learning to existing knowledge is a common factor of when elaboration will be useful (Endres et al., 2017; Larsen et al., 2013), consistent with earlier work that showed that elaborate study techniques are more beneficial when concepts are novel (Willoughby et al., 1994).

Experiment 7 specifically looked at whether elaborate retrieval that was not required for final test accuracy, as the information that was elaborately retrieved was not tested,

would show a larger testing effect. Results suggest that elaborating during retrieval, even when elaborating did not directly include the desired answer, was useful to the testing effect. Experiment 7 found that when restudy equivalent practice tasks are given, the benefit of retrieval practice is greater. Therefore, in line with points discussed previously in this chapter, the broader testing effect literature, in seeking answers to the mechanisms of the testing effect should manage to control the restudy equivalent such that results cannot be explained as a practice benefit.

As outlined in the introduction, much of the work looking at broader processing of the originally studied items during the retrieval practice task, has demonstrated this to be a useful learning strategy to achieve transfer of knowledge from retrieval practice (Butler, 2010; Hinze et al., 2013; Pan & Rickard, 2018). However, there has been limited work to date directly assessing the contribution of greater processing during retrieval practice, under the design conditions employed herein. Meaning this area is yet to be explored with the appropriate restudy control (Hinze et al., 2013) and without multiple opportunities for feedback or restudy in the retrieval practice condition (Butler, 2010). Therefore whilst the aim of this chapter's experiments is to assess meaning being achieved during the retrieval practice task, experiment 8 will additionally address whether the present findings will be extended to achieve wider meaning over the studied materials by including a simple manipulation of transfer knowledge.

One issue to note is that the extent of elaboration achieved during the retrieval practice task might have been limited due to the time limit imposed. Therefore for the additional experiments with this design given in this chapter (experiments 8 and 9), the time limit for retrieval practice trials will be removed.

4.3 Experiment 8

The aim of this experiment was to examine whether more meaningful retrieval practice could influence the transfer testing effect. For studies of the testing effect that have examined some form of transfer of knowledge between the retrieval practice task and the final test there is evidence to suggest that is a useful learning strategy. For ex-

ample, Butler (2010) showed that cued retrieval practice that consisted of elaborate short answers benefited both fact retrieval and inference retrieval at final test. Yet elsewhere, limited transfer knowledge has been found following basic fact retrieval during retrieval practice (Pan & Rickard, 2017), when the retrieval practice task and final practice task were multiple choice tests. No work has yet examined a direct manipulation of the amount of meaningful processing achieved during the retrieval practice task and the magnitude of the transfer effects. Experiments 8 and 9 explored this issue. This was achieved by examining whether elaborate retrieval processes that occurred at the same time as fact retrieval helped consolidate information across the fact more broadly.

In order to be able to compare these results as realistically to the initial results from experiment 7, the final test questions and answers were kept the same and the retrieval practice item was changed. Experiment 8 looked at whether practising retrieval for one noun in the sentence with additional elaboration, could transfer learning to retrieval of a different noun in the original fact at final test.

Based on previous work the hypothesis was that more elaborate retrieval would lead to a greater transfer effect. However, additional accounts such as the episodic context account and the effortful processing explanations would also predict that the elaborate retrieval task should confer a greater benefit on the final test, although for different reason. To date these perspectives have not been formally tested in relation to the transfer testing effect. Both experiments 8 and 9 assess the extent to which more meaningful processing leads to more meaningful learning, in the form of transfer knowledge.

4.3.1 Method

Participants and Design

Participants were 62 psychology students at Plymouth University. Participants received course credit or were paid for their time at £2.50 per fifteen minutes, or in course credit if they were registered students of psychology at the university and taking part during term time. Participants were aged between 18 and 62 years (M = 22.47, SD =

6.87), 79% female. Paid participants were paid in Amazon vouchers via email, £10 for completing both parts of the study. Twenty one participants were paid for taking part, 7 were test and 14 were restudy.

Experiment 8 utilised a 2 (practice task; restudy, test) x 2 (practice type; *what*, *what & why*) mixed design, with learning exercise as a between-subjects factor and practice type as a within-subjects factor. The order in which the elaborate or fact practice trials were presented was counterbalanced by participant.

Sample Size Calculation

The calculation for the sample required for this particular experiment was conducted in G*Power. As there was no reliable marker for the size of the effect to expect. The sample size was designed to detect a medium-sized main effect (Rowland, 2014), cued recall effect sizes tended larger than this g = .61. The G*Power (Faul et al., 2009) analysis was based on an ANOVA analysis with two between group factors and two within group factors. This gave each group size sample of 24.5 participants, with power of 0.80. Each group was rounded up to 32 per condition, to compensate for any additional variance associated with online platform. As the current study required 2 within-subjects groups, the final sample was 64 participants. Due to the time pressures associated with using the university participation pool at the end of the semester, the sample size achieved was 31 participants per condition, 62 in total.

Materials

The materials were almost identical to the materials used in experiment seven with the following exceptions. Instead of a noun object towards the end of the sentence being retrieved during the practice task, a noun object located elsewhere in the sentence was the target for test practice. The final test consisted of the same test questions as experiment 7. The full details of the study materials for experiment 8 can be found in appendix C.2.

Procedure

The procedure was almost identical to the procedure in experiment 7. However, the Gorilla Online Experiment builder (www.gorilla.sc) was used to create and host the experiment (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) because of the need to conduct socially-distanced testing during the Covid-19 pandemic. Data was collected between 22 March 2020 and 11 May 2020 and participants were recruited through the University of Plymouth SONA participant pool management software. Allocation to condition was counterbalanced. Participants were instructed to complete the session when they could do so one sitting and free of distractions. Besides this participants received the same instructions as the previous experiment. Participants completed a one minute break in between each phase of the study, in which they completed a forced choice arithmetic quiz. At the end of the first session participants were informed that they would be sent a link to the second part of the study in 48 hours. They completed the second session on average 2.95 days later than the first session.

4.3.2 Results

All analyses were computed in JASP (JASP Team, 2020) for both frequentist and equivalent bayesian tests (given where appropriate). Descriptive statistics for the main effects of interest are given in table 4.2.

Coding Responses

A new set of coding criteria was made for the retrieval practice test for the new noun *what* targets. The coding specifications were similar to experiment 7, whereby a number of synonym targets were accepted as well as spelling variations. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible.

Delay Period

Due to the fact that the experiment was conducted online, some control over the delay period by which participants completed the second part of the study was lost and there-

fore a check to see whether the restudy and retrieval practice participants completed the second part at a similar time was completed. A Mann-Whitney test (due to violation of normality) revealed no evidence for differences in the delay between the restudy practice participants (*Mdn*=2.47 days) and test practice participants (*Mdn*=2.67 days), *U*=458, *z* = -0.05, *p* = .76, *r*= -.05.

Ratings and Response Times

Two sets of analyses were conducted on the ratings for each animal. Independent samples t-tests were conducted between the restudy and test group for their average ratings prior to seeing the study materials, for both familiarity and interest. This revealed similar ratings in familiarity between the restudy practice group (*M*=2.88, *SD*=0.74) and the retrieval practice group (*M*=2.67, *SD*=0.92) prior to seeing the study materials, *t*(60)=1.01, *p*=.32, *d*= 0.26, BF_{10} = 0.40. This pattern was also seen for the interest ratings, where no evidence of differences between the restudy practice (*M*=3.74, *SD*=0.72) and retrieval practice groups' ratings (*M*=3.53, *SD*=0.71) could be seen, *t*(60)=1.18, *p*=.24, *d*= 0.30, BF_{10} = 0.47.

In addition, further paired samples t-tests were conducted across the whole sample for average familiarity scores before and after the study phase, which revealed that participants rated their familiarity as lower before studying the materials (*M*=2.77, *SD*=0.84) compared with after studying the materials (*M*=1.98, *SD*= 0.93), *t*(61)=6.22, $p<.001$, *d*= 0.79, $BF_{10} > 150$ (267711.63). Once more suggesting participants were not already familiar with the information they were studying. Whereas participants rated their interest as lower prior to studying the materials (*M*=3.63, *SD*=0.72), compared to after studying the materials (*M*=3.98, *SD*=0.85), *t*(61)=-3.61, $p<.001$, *d*=-0.46, BF_{10} = 40.34. Suggesting the information was of interest to participants.

Main Analyses

Initial test performance. Initial test performance was compared using a paired samples t-test, between the two types of practice trials. Results revealed the same

Table 4.2
Initial and Final Test Accuracy in Experiment 8 as a Function of Practice Task and Practice Type

	Initial test		Avg	Final test		Avg
Practice task	What	W & why	IT	What	W & why	FT
Restudy	n/a	n/a	n/a	.69 (0.17)	.65 (0.22)	.67 (0.18)
Test	.74 (0.17)	.73 (0.17)	.73 (0.16)	.68 (0.18)	.66 (0.21)	.67 (0.16)

Note. The values represents mean percentages of target words recalled, SDs given in parentheses.

accuracy rates for retrieving the target *what* element in both the *what* practice trials (*M*=.74, *SD*=0.17) and the *what & why* practice trials (*M*=.73, *SD*=0.17), *t*(30)=0.78, *p*=.44, *d*=0.14, BF_{10} =0.25.

Final test performance. A 2 (practice type; restudy, test) x 2 (practice type; *what*, *what & why*) mixed ANOVA was conducted, with practice task as a between-subjects factor and practice type as a within-subjects factor, to examine the impact of practice type on the transfer testing effect. Results revealed no main effect of practice task, with the restudy practice group (*M*=.67, *SD*=0.18) performing at the same level to the test practice group (*M*=.67, *SD*=0.16), *F*(1,60) = 0.002, $p < .97$, $\eta_p^2 < .001, BF_{10}$ = 0.21. There was no main effect of practice type, with the *what* practice trials (*M*=.68, *SD*=0.18) showing the same level of final test accuracy to *what & why* practice trials (*M*=.66, *SD*=0.21), *F*(1,60) = 0.98, *p* = .33, $\eta_p^2 = .02, BF_{10}$ = 0.21. In addition there was no evidence for an interaction between these two factors, *F*(1,60) = 0.32, *p* = .57, $\eta_p^2 < .01, BF_{10}$ = 0.29. Results of the final test are depicted in figure 4.2.

4.3.3 Discussion

Results revealed no evidence of a transfer testing effect and no evidence that the magnitude of the transfer testing effect is altered by the amount of meaningful processing achieved during the initial retrieval practice task. There are a number of points worth noting here. Firstly, there is the possibility that due to fewer controls being imposed

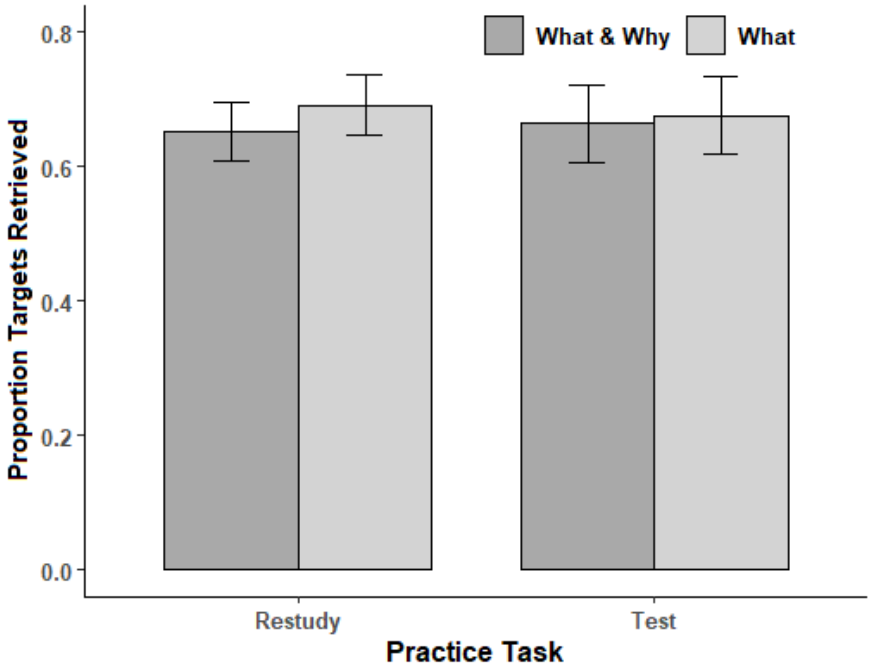

Figure 4.2. Mean target retrieval at final test as a function of practice task and practice type in experiment 8. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

in the running of the task that these results are not directly comparable to experiment 7 results. Indeed, the restudy group performed at far superior levels to the results for the same group in experiment 7. It could have been that as the online experiment was taken at times that suited the participant, that superior levels of focus were achieved for the restudy group that brought their performance up to the levels of the test group. This might be a useful area for future work to assess, perhaps with additional assessment of the participants decisions in taking part in an online experiment at a particular time of day.

It is also possible, that restudy participants could have used additional prompting, like a pen and paper to write down notes despite not being instructed to. However, the restudy group did have fixed and limited presentation time, therefore it seems unlikely, due to the fact that it would seem additionally effortful to write the information down at speed and we would expect higher levels of accuracy if notes were taken. In this experiment accuracy was comparable to accuracy in experiment 7.

Despite the lack of benefit to the transfer testing effect reported here, these results could be seen to be consistent with previous work, which has shown that elaborate retrieval can be more beneficial for comprehension rather than detail learning (Endres et al., 2017). As experiment 8 assessed detail learning, it is perhaps not surprising that transfer to another target where cues have not been as extensively formed has not been made. The results could suggest that specific links could be made when items are elaborated on in relation to a specific target. The current findings are not consistent with the results of Pan and Rickard (2017) who found that there was support for more extensive cues leading to limited transfer of target information. However, the items in the previous work (Pan & Rickard, 2017) were directly related, therefore it might be necessary to boost the relatedness between the items in order to achieve this transfer.

Therefore experiment 9 will examine whether encouraging further relational processing during retrieval of the *what* element of the fact will further encourage basic transfer on the final test.

Based on previous findings that increased cues and relational processing lead to a greater likelihood of transfer advantage, the prediction is that transfer knowledge will be shown to a greater extent in the meaningful processing condition.

4.4 Experiment 9

In order to encourage greater relational processing during retrieval practice of the *what* element, the object to be retrieved was changed from a noun object in experiment eight to a verb object in experiment nine. Previous work has indicated that processing verb objects can be a more difficult task (Gentner & France, 1988; Gillette, Gleitman, Gleitman, & Lederer, 2000), that verbs encourage greater processing of the meaning of a sentence (Kintsch, 1974) and that they encourage greater processing of the sentence context (Barclay, Bransford, Franks, McCarrell, & Nitsch, 1974). They can further help to organise and activate existing knowledge (Ferretti, McRae, & Hatherell, 2001). This suggests that verb objects serve to add meaning to learning materials. Therefore, including a verb object as the retrieval practice item may boost the overall relatedness

achieved in the *what* element and enable meaningful processing during the *what & why* element achieved during retrieval practice, that will lead more successfully to transfer learning.

4.4.1 Method

Participants and Design

Participants were 63 psychology students at Plymouth University. Participants received course credit or were paid for their time at £2.50 per fifteen minutes, or in course credit if they were registered students of psychology at the university and taking part during term time. Participants were aged between 18 and 52 years (M = 21.98, SD = 6.70), 90% female. Paid participants were paid in Amazon vouchers via email, £10 for completing both parts of the study and £5 if for some reason they were only able to complete the first part. Sixteen participants were paid for completing the study, 4 were test participants and 12 were restudy participants.

Experiment 9 utilised a 2 (practice task; restudy, test) x 2 (practice type; *what*, *what & why*) mixed design, with practice task as a between-subjects factor and practice type as a within-subjects factor.

Sample Size Calculation

The calculation for the sample required for this particular experiment was conducted in G*Power. As there was no reliable marker for the size of the effect to expect. The sample size was designed to detect a medium-sized main effect (Rowland, 2014), cued recall effect sizes tended larger than this g = .61. The G*Power (Faul et al., 2009) analysis was based on an ANOVA analysis with two between group factors and two within group factors. This gave each group size sample of 24.5 participants, with power of 0.80. Each group was rounded up to 32 per condition, to compensate for any additional variance associated with online platform. As the current study required 2 within-subjects groups, the final sample was 64 participants. Due to restrictions in time collecting the data from this participant pool the final sample was 63.

Materials

The materials were almost identical to the materials used in experiment 8 with the following exceptions. Instead of a noun object located elsewhere in the sentence being the target for retrieval practice, the object to retrieve was a verb. The final test consisted of the same items as experiments 7 and 8. The full details of the practice test questions for experiment 9 can be found in appendix C.2.

Procedure

The procedure was identical to the procedure in experiment 8, also using the Gorilla Online Experiment builder (www.gorilla.sc) to create and host the experiment (Anwyl-Irvine et al., 2020). Data was collected between 22 March 2020 and 11 May 2020 and participants were recruited through the University of Plymouth SONA participant pool management software. This was run during the same period as experiment 8 and participants were only able to take part in one of the experiments. They were randomly assigned to one of the conditions across the two experiments upon sign-up. Allocation to condition was counterbalanced. As with experiment 8 participants were instructed to complete the session when they could do so in one sitting and free of distractions. Besides this participants received the same task instructions as experiments 7 and 8. Participants completed a one minute break in between each phase of the study, in which they completed a forced choice arithmetic quiz. At the end of the first session participants were informed that they would be sent a link to the second part of the study in 48 hours. The second session on average was completed 2.88 days after the first session.

4.4.2 Results

All analyses were computed in JASP (JASP Team, 2020) for both frequentist and equivalent bayesian tests (given where appropriate). Descriptive statistics for the main effects of interest are given in table 4.3.

Coding Responses

A new set of coding criteria was made for the practice test for the new verb targets. The coding specifications were the same as previous experiments, with a number of synonyms of correct targets as well as spelling variations where another word was not spelled. Intrusions were classified as such, however as the number of intrusions were negligible no formal analysis was possible.

Delay Period

Due to the fact that the experiment was conducted online, some control over the delay period by which participants completed the second part of the study was lost and therefore a check to see whether the restudy and test participants completed the second part at a similar time was completed. A Mann-Whitney (due to violation of normality) revealed no differences in this delay for the restudy practice (*Mdn*=2.39 days) and test practice participants (*Mdn*=2.21 days), U=512, z = 0.02, p = .83, r= .03.

Ratings and Response Times

Two sets of analyses were conducted on the ratings for each animal. Independent samples t-tests were conducted between the restudy and test group for their average ratings prior to seeing the study materials, for both familiarity and interest. This revealed the same levels of familiarity between the restudy practice group (M=2.67, SD=0.88) and the test practice group (M=2.54, SD=0.81) prior to seeing the study materials, $t(61)$=0.58, p=.56, d=.15, BF_{10} = 0.30. This pattern was also seen for the interest ratings for restudy practice (M=3.48, SD=0.80) and test practice (M=3.45, SD=0.52), $t(60)$=0.21, p=.81, d=.05, BF_{10} = 0.26.

In addition, further Wilcoxon signed-rank tests were conducted (due violation of normality in the samples) across the whole sample for average familiarity scores (between 1 and 5) before and after the study phase, which revealed that participants' familiarity was rated as higher before studying the materials (*Mdn*=2.67) compared with after studying the materials (*Mdn*=1.67), T=1530, $p<.001$, r=.52, $BF_{10} > 150$ (1.886e+9).

Table 4.3
Initial and Final Test Accuracy in Experiment 9 as a Function of Practice Task and Practice Type

	Initial test		Avg	Final test		Avg
Practice task	What	W & why	IT	What	W & why	FT
Restudy	n/a	n/a	n/a	.70 (0.18)	.69 (0.19)	.70 (0.14)
Test	.75 (0.13)	.74 (0.16)	.75 (0.12)	.68 (0.16)	.71 (0.18)	.69 (0.14)

Note. The values represents mean percentages of target words recalled, SDs given in parentheses.

Suggesting that participants were not overly familiar with the study materials. Whereas participants' interest ratings were lower prior to studying the materials (*Mdn*=3.67), compared to after studying the materials (*Mdn*=4.33), *T*=106, $p<.001$, *r*=-0.95, $BF_{10} > 150$ (4.598e+7). Suggesting that participants' interest did not decrease after studying the materials.

In experiments 8 and 9, participants were able to direct their own practice test timing. To assess the relative difficulty of the different tasks an independent samples t-test was run on the overall response times to retrieve the target item in the *what* element of the fact. Results revealed that during the practice task, the time taken to retrieve the the noun object in experiment 8 (*M*=6679 ms, *SD*=1614) was shorter than the verb object in experiment 9 (*M*=8290 ms, *SD*=1979), $t(61) = -3.53$, $p < .001$, $d = -0.89$, $BF_{10} = 37.67$.

Main Analyses

Initial test performance. Initial test performance was compared using a paired samples t-test, between the two types of practice trials. The results showed the same level of recall for *what* practice trials (*M*=.75, *SD*=0.13) and the *what & why* practice trials (*M*=.74, *SD*=0.16), $t(31)=0.12$, $p=.91$, $d=0.14$, $BF_{10}=0.19$.

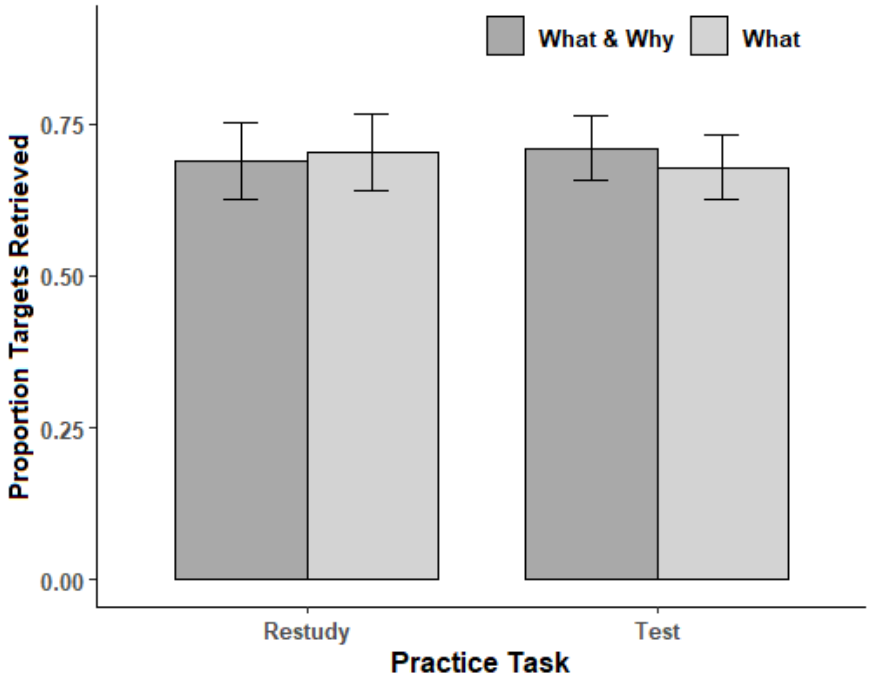

Figure 4.3. Mean target retrieval at final test as a function of practice task and practice type in experiment 9. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

Final test performance. A 2 (practice type; restudy, test) x 2 (practice type; *what*, *what & why*) mixed ANOVA was conducted, with practice task as a between-subjects factor and practice type as a within-subjects factor, to examine the impact of the practice type on the transfer testing effect. Results revealed no main effect of practice type, with the restudy practice group (*M*=.70, *SD*=0.14) overall performing the same as the test practice group (*M*=.69, *SD*=0.14), $F(1,61) = 0.003$, $p < .96$, $\eta_p^2 < .001, BF_{10} = 0.17$. There was no main effect of practice type, with the *what* practice trials (*M*=.69, *SD*=0.17) showing similar levels of final test accuracy to *what & why* practice trials (*M*=.70, *SD*=0.18), $F(1,61) = 0.07$, $p = .79$, $\eta_p^2 = .001, BF_{10} = 0.14$. In addition there was no evidence for an interaction between these two factors, $F(1,61) = 0.70$, $p = .41$, $\eta_p^2 = .01, BF_{10} = 0.34$. Results approached conclusive evidence of no evidence, based on the bayes factor. Results of the final test are depicted in figure 4.3.

4.4.3 Discussion

The results of experiment 9 again showed no evidence for an interaction between practice type and practice format.

The manipulation in experiment 9 attempted to increase processing of the relatedness of the *what* element of the fact by requiring the retrieval of a verb instead of a noun, which is thought to increase organisation and activation of existing knowledge (Ferretti et al., 2001) and comprehension of information (Kintsch, 1974). However, it might be that comprehension of materials was already sufficiently high, that any boost to processing in this regard was no longer useful to retrieval or transfer. The current findings suggest that retrieval practice does not appear to be at a deficit in comparison to restudy, for transfer effects with highly related information, somewhat consistent with previous work (Pan, Wong, Potter, Mejia, & Rickard, 2016), whereby highly associated information benefits from feedback when the answer requires transfer to another target.

It is interesting to note here, that the practice accuracy is again higher than in experiment 7 for both restudy and test practice. It is possible that the study materials appealed to participants to a greater extent during the climate of a coronavirus lockdown, encouraging maximum engagement and limited the deficit to the restudy groups that we would typically expect to see in these tasks and saw in experiment 7.

It is also worth noting here that experiments 7, 8 and 9 offer a particularly strict assessment of the contribution of meaningful retrieval practice, as in each case the additional meaningful retrieval did not directly include the desired answer for the final test. Therefore the effects reported here are likely to be an underestimate of the relevance of this component in the wider literature, as elaborating directly with items that need to be retrieved at the final test is conceivably a more useful exercise. However, as clearly shown here, these current results do not demonstrate a unique benefit over restudy, however, future work might look to change components of the materials, like difficulty or cohesiveness (Roelle & Nückles, 2019), to assess the limits of these results.

4.5 Experiment 10

The results so far indicate that meaningful processing might be useful to retention and the testing effect, but possibly not transfer to the same extent or with highly related materials. The results of experiment 7 can be interpreted as support for either elaborate retrieval or context reinstatement as the reason for increased benefit to testing for the *what & why* trials. Indeed research to date suggests that both of these elements are likely to be useful (Karpicke et al., 2014). The materials utilised in the previous experiments of this chapter were easy to elaborate on and previous work has suggested that putting information into the learner's own words could be key to any benefit (Endres et al., 2017; Hinze et al., 2013; Larsen et al., 2013). Yet there are many other material types that are seen throughout educational practice and in some cases the possibility to put the materials into one's own words is limited. Study materials that depict a process are ubiquitous in sciences and technologies, yield large testing effects when feedback is utilised (Karpicke & Blunt, 2011) and when the retrieval practice task enables the learner to construct their own knowledge during the retrieval task (Blunt & Karpicke, 2014). However, these effects have not been directly assessed in the absence of feedback or when the opportunity to construct knowledge is limited.

Experiment 10 will seek to explore whether the results of experiment 7 will be replicated when the more meaningful practice trials enable reinstatement of the previous context, but a reduced ability to elaborate or construct one's own knowledge. Using a process text, the information is still highly interrelated and meaningful, but the opportunity to reconstruct knowledge will be reduced to retrieval of specific missing items. Instead of offering participants the opportunity to retrieve and reinstate a large portion of the previous study item, all of the item will be reinstated, but additional retrieval will be constrained to an item in the additional information. Experiment 10 utilised a new set of learning materials. The new materials required participants to study pairs of facts about coffee production that were presented in the chronological order of the process. The nature of the meaningful processing task was limited to the retrieval of an additional item contained in the second item of the pair.

Based on the work of the episodic context account (Karpicke et al., 2014), the trials that enabled participants to reinstate all of the previous context should result in a larger testing effect than trials that did not.

4.5.1 Methods

Participants and Design

Participants were 56 students and members of the public that took part in the study on campus at Plymouth University. Participants were paid for their time at £2 per fifteen minutes, or in course credit if they were registered students of psychology at the university and taking part during term time. Participants were aged between 18 and 49 years (M = 21.7, SD = 6.24), 84% female.

The design for experiment 10 was very similar to the previous experiments in this chapter, there was a 2 (practice task; restudy, test) x 2 (practice format; *what*, *what x2*) design, with practice task as a between-subjects factor and practice type as a within-subjects factor.

Sample Size Calculation

The calculation for the sample required for this particular experiment was conducted in G*Power. As there was no reliable marker for the size of the effect to expect. The sample size was designed to detect a medium-sized main effect (Rowland, 2014), cued recall effect sizes tended larger than this g = .61. The G*Power (Faul et al., 2009) analysis was based on an ANOVA analysis with two between group factors and two within group factors. This gave each group size sample of 24.5 participants, with power of 0.80. Each group was rounded up to 28 participants. As the current study required 2 within-subjects groups, the final sample was 56 participants.

Materials

Materials were generated about the production of coffee. The information was taken from a series of websites. These materials were designed to more closely mimic the types of materials that participants could encounter as part of a course. The structure

of the materials was also set up to mimic the type of revision that participants might engage with, typically students will choose to chunk large sections of materials into smaller chunks to revise. Therefore this seemed to be a useful way to break up the learning materials in a way that might mimic a real learning scenario. The materials were pairs of facts that complemented each other, but were not explanatory of one another like in experiments 7, 8 and 9. The presentation of the facts was in sentence form. Two sentences were listed together as one item to study. For retrieval practice, noun objects in each of the sentences for one study item were removed and replaced with the initial letter cue stem. Each of the items to retrieve were unique items. For the *what* test practice trials only the initial sentence with one missing word was shown. Whereas, for the *what x 2* test practice trials, both sentences were shown with two words missing and the cue stems visible in their place. For the final test, the test items were identical to the *what* test practice trials. Therefore all participants had seen all items in this form, but had also seen half of the items accompanied by the additional item that it was studied with. The full details of the study materials for experiment 10 can be found in appendix C.3.

Procedure

Participants were recruited to take part through the University of Plymouth SONA participant pool management software. Participants signed up to take part in both parts of the experiment. The first session took 30 minutes and the second session took around 15 minutes. Participants attended the lab and were tested in groups of up to six people. For each experimental session, participants were sat at a partitioned desk with their own PC. Participants wore headphones throughout session one. Besides the filler task, all elements of the experimental task were presented in PsychoPy2 (Peirce et al., 2019). Participants were told that the first session would be a learning session and that the second session would be a test session. Participants were told that they would learn some information about the production of coffee in three different parts and that they would have two learning opportunities for each part, before moving on to the

next part.

Participants started the study phase of the first part of coffee production and studied the pairs of items for a set time of 15 seconds per pair. Once all 18 items in the first part had been studied once, participants completed a short filler task where they completed a number search or sudoko puzzle; both puzzles were supplied on a double-sided sheet of paper. A numerical filler task was chosen for the break, to minimise the likelihood of participants rehearsing something associated to the studied materials. Participants then moved on to the practice phase for part one. Participants were assigned to one of two practice phases for the entire experiment, restudy or test practice.

Participants in the restudy condition were able to restudy the items again in the same order as before, but this time they were able to move their learning on at their own pace by pressing the space bar to move on. Half of the items were presented in the same manner as before, with the pair of facts presented on screen and half were presented as only the first item from the pair. Participants were told to only restudy the item being presented.

Participants in the test practice condition also had the items presented in the same order as they had previously studied them. Half of the items included the full pair of items, each with a word missing, *what x 2* trials. Participants saw a cue letter at the start of each of the missing words. Participants had to retrieve the items in the order that they had been presented, filling in the missing word for the first items before the second item. Half of the items only had the initial item in the pair presented again, this time with a word missing in the *what* trials. Again, participants were given the letter stem and asked to retrieve the word they had previously studied. Participants could answer at their own pace. Participants had a maximum of 45 seconds in both practice conditions for each trial, after this time they were moved on to the next trial.

After participants had finished the practice phase for the first part they completed a one minute filler task and proceeded in the same manner to complete the remaining two parts. Following completion of all three parts, participants were reminded of the

second session time, thanked and dismissed. Participants returned to take part in the test session after 48 hours. For the test session participants were told that they would see a single fill-in-the-blank test item for each of the items they studied in the first session. The items in the final test were made up of only the *what* items associated with the first fact. The *what* items from the second fact were not tested on the final test. Participants were told to try to recall the item as best they could. The second session took around 15 minutes, following this participants were debriefed and dismissed.

4.5.2 Results

All analyses were computed in JASP (JASP Team, 2020) for both frequentist and equivalent bayesian tests (given where appropriate). Descriptive statistics for the main effects of interest are given in table 4.4.

Coding Responses

Answers were coded in the same way to previous experiments. Items were coded blind to condition. Plurals incorrectly present or absent, obvious spelling mistakes and two letter changes to make up correct words (but not another word) were coded as correct. Due to the inclusion of cue letter stems during testing in experiment 10 very few intrusions were registered in participant's answers, therefore no formal analysis of intrusions was possible.

Response Times

Two paired samples t-tests were computed to assess the response times taken to answer the practice test questions in the test condition. One test was computed between the response time to the initial test item between the *what* trials that required one word to be inputted and *what x 2* trials that required two words to be inputted. Results showed that response times for the initial answer in the *what x 2* trials were longer (M = 9.96, SD = 2.81) than the response times to the initial answer in the *what* trials (M = 8.63, SD = 2.24), $t(27) = 3.83$, $p < .001$, $d = 0.72$, $BF_{10} = 48.82$. This could have been due to both FITB items being on screen at the same time. In addition, a comparison

Table 4.4
Initial and Final Test Accuracy in Experiment 10 as a Function of Practice Task and Practice Type

	Initial test		Avg	Final test		Avg
Practice task	What	What x 2	IT	What	What x 2	FT
Restudy	n/a	n/a	n/a	.56 (0.16)	.54 (0.16)	.55 (0.14)
Test	.65 (0.13)	.64 (0.14)	.65 (0.12)	.63 (0.15)	.62 (0.15)	.63 (0.13)

Note. The values represents mean percentages of target words recalled, SDs given in parentheses.

was made between the overall trial times for the *what* trials and the *what x 2* trials. Here again the *what x 2* (M = 17.66, SD = 4.79) trials took longer overall to complete than the *what* trials (M = 9.37, SD = 2.44), $t(27)$ = 14.69, $p < .001$, d = 2.78, $BF_{10} >$ 150 (3.074e+11), although this was to be expected based on the additional item being retrieved in these trials.

Main Analyses

Initial test performance. Accuracy on the initial *what* item on the practice task was compared between the *what* and *what x 2* trials in a paired samples t-test. Results showed that performance during retrieval practice was equivalent between the *what x 2* trials (M = 0.64, SD = 0.14) and *what* trials (M = 0.65, SD = 0.13), $t(27)$ = -0.66, $p <$ .51, d = -0.13, BF_{10} = 0.25.

Final test performance. For the main analyses a 2 x 2 mixed ANOVA was computed, with practice task (restudy, test) as a between subjects factor and practice type (*what*, *what x 2*) as a within-subjects factor. Results revealed some evidence for a main effect of practice task which was not wholly supported by the bayesian analysis, with restudy practice (M = 0.55, SD = 0.14) performance on the final test being inferior to test practice performance on the final test (M = 0.63, SD = 0.13), $F(1, 54)$ = 3.97, p = .05, $\eta_p^2 = .07, BF_{10}$ = 1.62. There was no evidence for a main effect of practice

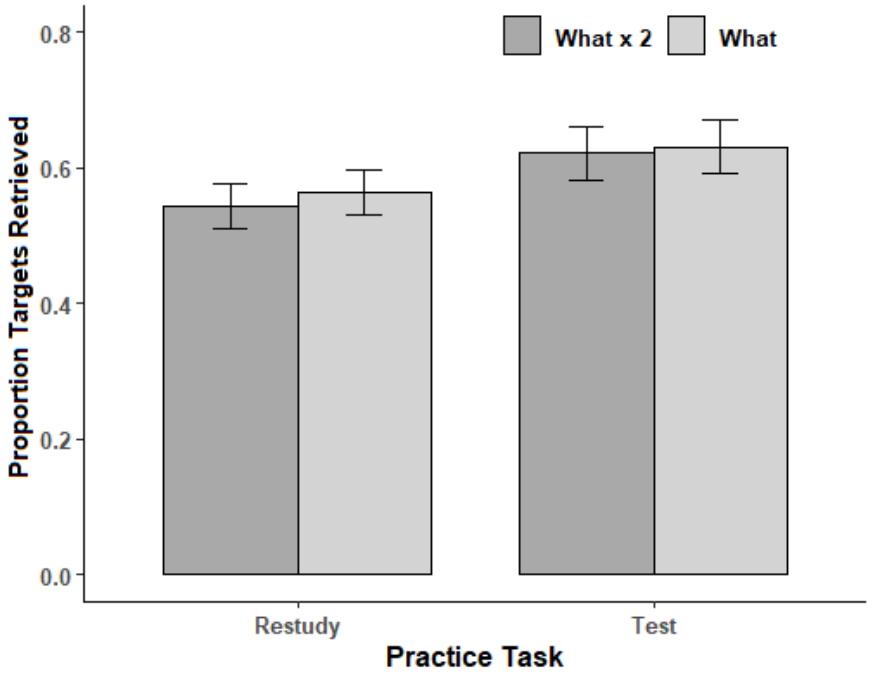

Figure 4.4. Mean target retrieval at final test as a function of practice task and practice type in experiment 10. Error bars depict the standard error of the mean with adjusted calculations for within-subjects designs as given in Morey et al. (2008).

type, with *what x 2* trials overall showing similar performance to *what* trials, $F(1, 54) = 0.70$, $p = .41$, $\eta_p^2 = .01, BF_{10} = 0.21$. Further there was no evidence for an interaction between practice task and practice type, $F(1, 43) = 0.01$, $p = .74$, $\eta_p^2 < .01, BF_{10} = 0.14$. The results of the final test are depicted in figure 4.4.

4.5.3 Discussion

The results of experiment 10 showed evidence of a testing effect, albeit a small effect. There was no evidence for an interaction between the practice task and the practice type. As the changes to the materials impeded elaborate processing or knowledge construction during retrieval, this could be the reason that no interaction was seen. There was no significant difference seen between the restudy trial types or the retrieval practice trial types. This suggests that the changes in this study design impacted both processes seen in experiment 7, namely the increase to retrieval accuracy in the elaborate trials and the decrease to restudy accuracy. However, the numerical trend for the restudy accuracy to be better in the *what* only restudy trials is reminiscent of the

findings from experiment 7.

Experiment 10 was primarily designed to assess whether reinstating the context in full while also enabling a retrieval opportunity would be more beneficial than not reinstating the context in full during the practice task, when the restudy opportunity was matched and opportunity to reconstruct knowledge was impaired. Results showed that the increased context reinstatement was not more beneficial, there could be a number of reasons for this. The reason the additional context reinstatement available during additional retrieval practice was not more beneficial could be related to the fact that context reinstatement could also rely on context reconstruction. Participants were not required to reconstruct the previous information as they had in experiments 7-9, but they had an opportunity to re-experience the full study item at the same time as retrieving key information. This could therefore suggest that reinstating the previous context is only beneficial to retrieval when participants also reconstruct the previous information in full through retrieval, as we see in free recall memory tests (Blunt & Karpicke, 2014; Karpicke & Blunt, 2011). This idea is consistent with the results from the transfer literature also (Hinze et al., 2013).

It is also possible that the retrieval practice task that required greater context reinstatement/elaboration in experiment 7, also involved more related information. Taken together with the results from experiment 10, this suggests that additional retrieval cues being actively retrieved is beneficial to specific retrieval in line with the elaborate retrieval hypothesis (Carpenter, 2009) and episodic context account (Karpicke et al., 2014). However, based on the current findings, it is not possible to deduce whether it is more important if these are semantic or contextual in nature, or whether this depends on the nature of the materials being retrieved. Recent work has suggested that contextual information is useful for retrieval (Schwoebel et al., 2018), yet again work has not fully controlled for the restudy task.

What is encouraging from the current work and this echos previous research (Endres et al., 2017), is that retrieving additional information is not detrimental to learning specific information.

Further work should look to explore the role of generating cues, context reinstatement and relational processing during the retrieval practice task, and assessing how these factors contribute to the testing effect.

4.6 General discussion: Experiments 7-10

The results described in this chapter make for a slightly confusing picture. In experiment 7, there was a main effect of testing and an interaction with the practice task employed and the meaningful processing manipulation. Results showed that a more meaningful retrieval practice task, which required participants to recall more of the information given in the study fact boosted the benefit of testing over a restudy equivalent task. Experiment 8 looked to examine whether the benefit associated with increased recall in experiment 7 could extend to a simple transfer task, from one noun object retrieved during the practice task, to a new noun object in the final test. Results did not show a main effect of practice task or an interaction with the meaningful processing manipulation. This was somewhat surprising, but could have been due to the conditions of the experiment being conducted online during the coronavirus, boosting interest levels.

Experiment 9 examined whether a transfer effect would been shown when a verb object was retrieved during retrieval practice, due to the fact that verbs can assist in meaningful processing and comprehension (Ferretti et al., 2001; Kintsch, 1974), which could be a feature where previous transfer effects have been found (Pan & Rickard, 2017). Results were consistent with experiment 8, showing high accuracy across conditions, possibly suggesting that increasing the amount of information retrieved does not impede transfer which is again consistent with previous findings (Butler, 2010; Hinze et al., 2013; Pan & Rickard, 2018). Experiment 10 looked to explore whether the results of experiment 7 would replicate, when more meaningful processing involved reinstating context during retrieval, but not an ability to elaborate or construct knowledge. Results revealed a testing effect, however, the retrieval of the additional item during full context reinstatement of the study item did not boost the benefit of retrieval compared to a restudy control, which is slightly at odds with the episodic context account

(Karpicke et al., 2014).

The results from this chapter further serve to highlight how the nature of the testing effect is not a static entity and is likely changeable with the many different factors that are being explored. To this end, the current research highlights the benefit of matching the restudy practice task as much as possible to the specific retrieval practice task. This way the effects that are uniquely a factor of retrieval practice as opposed to any equivalent task that strengthens memory can be more readily determined.

The results of experiment 9 demonstrated the largest accuracy results for the retrieval practice condition and demonstrated this under conditions of transfer. Therefore, this further suggests that additional relational processing, or contextual processing associated with increased comprehension, like increased processing of verb objects, is a useful strategy for studying materials. This could be a promising area for future work, as the current work suggests that properties associated with both relational processing and context reinstatement could be equally likely to benefit testing when they occur during retrieval practice.

The results further support the idea that increased processing during retrieval does not impede the efficacy of testing. Previous suggestions that increased elaboration during retrieval would encourage cue overload (Karpicke et al., 2014; Karpicke & Smith, 2012; Lehman & Karpicke, 2016) and would likely be a hindrance to testing have not found support in results from this chapter. In addition, as experiment 10 demonstrated a testing effect, but no benefit of multiple item retrieval, suggests once more that difficulty alone does not benefit retrieval when the restudy task is matched.

The nature of the designs in the current chapter have not included feedback. As previous studies have also shown that feedback can be useful for both specific (Rowland, 2014) and transfer learning (Butler, 2010; van Eersel et al., 2016), future work should look to isolate when feedback it most useful in this regard.

The series of four experiments presented in chapter four further highlights that the testing effect can be elusive. Here, the failure to observe a testing effect might be due to the fact that experiments 8 and 9 included transfer learning. Although transfer

effects have been shown previously, reliable effects are often found when the nature of the retrieval practice task is quite removed from the nature of the restudy task, whereby any processing differences might be maximised (Pan & Rickard, 2018). However, the failure could have been due to the time and conditions under which experiments 8 and 9 were run, for example, during the coronavirus lockdown on an online platform, which could have influenced results.

Chapter 5

Discussion

The current chapter will discuss the results of the previous three experimental chapters and put them into focus in light of the perspectives outlined in chapter one. Firstly, a summary of results is presented, followed by discussion of the null findings in this thesis. Further, results are discussed in line with the current thinking in the testing effect literature and recent work.

5.1 Meaningful Processing Explored

Across ten experiments the contribution of meaningful processing to the testing effect was explored. In chapters two and three, meaningful processing was manipulated within the study materials and in chapter four meaningful processing was manipulated during the practice task. In chapter two this took the form of revisiting the concepts that underpin the elaborate retrieval hypothesis, by replicating and attempting to extend the formative results associated with this perspective. The aim of this chapter was to explore the elaborate retrieval hypothesis in line with the testing effect phenomenon of the test-delay interaction (Adesope et al., 2017; Roediger & Karpicke, 2006b; Rowland, 2014). In chapter three, meaningful processing was further explored through the mediator effectiveness hypothesis and the structural coherence of the materials. In light of the consistent null results found in chapters two and three for the contribution of meaningful processing in the study materials to the testing effect, chapter four looked to explore meaningful processing in the retrieval task. This chapter will summarise the key findings of this thesis and further discuss how these results challenge current theory and practice.

5.2 Summary of Results

5.2.1 Chapter Two Results

In chapter two the contribution of meaningful processing of the study materials was explored in relation to the testing effect. This was achieved by revisiting a theory of the testing effect that has received some interest in the last decade, yet has not been extensively explored to date, the elaborate retrieval hypothesis. Chapter two more specifically examined whether the results from the elaborate retrieval hypothesis' original formulation (Carpenter, 2009), could be extended in line with the test-delay interaction found in the literature, whereby the size of the testing effect has been shown to increase when the delay to the final test increases (Adesope et al., 2017; Roediger & Karpicke, 2006b; Rowland, 2014). The original hypothesis suggested that increased elaboration occurring during the retrieval practice task for semantically weakly associated items, would lead to a benefit of retrieval practice. This was because weakly associated items required a longer search of memory and activated more related items during the search, that could serve as subsequent retrieval cues.

In experiment 1, Carpenter's results were almost totally replicated, giving confidence in the devised materials. There was indication that the weakly associated word pairs were more difficult to retrieve and were rated as less related than the strongly associated word pairs. In addition, the interaction between the initial and final test results found by Carpenter was replicated in experiment 1. This suggested that the weaker associates showed less forgetting than the stronger associates. However, when the final test type was changed in experiment 2 to match the cued recall test type at initial test, the findings from experiment 1 were not replicated. In addition, the final test testing effect analysis gave a negative testing effect. However, this was in line with results in the literature that have shown a benefit for restudy practice when the delay to final test is brief (Roediger & Karpicke, 2006b). The mechanistic property of this phenomenon has not been much given attention, due to the common interpretation that the restudy advantage with an immediate test results from all of the items being restudied and not

yet subject to forgetting, which retrieval practice protects against (Kornell et al., 2011; Wheeler et al., 2003). However, some attention has been paid to this phenomenon in free recall designs (Mulligan & Peterson, 2015; Peterson & Mulligan, 2013). Here, experimental design has been found to mitigate the negative testing effect, which is a restudy advantage found with free recall tests specifically. However, the assertion is that negative testing effects can occur in free recall tests due to serial order position information being better preserved for the restudy items than the retrieval items. This is because retrieval confers greater item-level encoding, as opposed to inter-item or relational encoding which restudy practice relies on for its advantage. As experiment 2 utilised a cued recall test with random item presentation, the negative testing effect results would sit somewhat at odds with this account, which predicts that retrieval should make better use of cue-target associations than restudy where serial order position is not informative for recall.

Experiment 3 added a delay of 3-5 days to the final test, rather than the original 5 minutes, and used the same cued recall final test from experiment 2. This time a large testing effect was found, but no evidence for an interaction was found between the meaningful processing manipulation (weak versus strong associates) and the testing effect (restudy versus test practice). The difference in response times between weak (longer RTs) and strong associates (shorter RTs) across all three experiments was consistent with additional memory search, elaboration, or increased difficulty being part of weak associate target retrieval. Therefore, while there was an indication that elaboration could be present during the retrieval practice task for the weaker associates, the only interaction found in the predicted direction was between test phase (initial versus final test) and the association strength (weak versus strong associates), or the test phase analysis in experiment 1. However, as this analysis did not represent an accurate forgetting distribution as the test types were not matched, this evidence is somewhat dubious.

For the remaining results, whereby initial and final tests were the same (experiments 2 & 3) and a comparison restudy practice task was included (testing effect

analysis experiments 1, 2 & 3), no evidence was found that more elaboration (for the weaker pairs) contributes to the testing effect. The weakly associated word pairs took longer to retrieve and were rated as less related, but did not show an increased benefit of testing at final test. Taken together these results suggested that Carpenter's original results were most likely due to the inclusion of a final free recall test and could not be extended in line with a test-delay interaction (experiment 3), whereby the final test was administered after several days and not minutes. These results will be discussed in more detail in relation to the elaborate retrieval hypothesis below. The results of chapter two led to a lack of confidence in the elaborate retrieval hypothesis as an explanation for the testing effect and the role of meaningful processing in the testing effect more broadly.

5.2.2 Chapter Three Results

Chapter three continued to look at whether aspects of meaningful processing in the study materials could yet show a contribution to the testing effect. Further theory and evidence from relevant research was explored, in relation to the role of mediating information during retrieval practice (experiment 4) and the structural coherence of the materials (experiments 5 and 6). Experiment 4 explored the mediator effectiveness hypothesis, which predicted that retrieval practice should increase the benefit associated with available mediating information, when compared to restudy and thereby impact the magnitude of the testing effect.

Experiment 4 manipulated the amount of mediating information available during the study phase. The stimuli were word pairs that consisted of an unusual English adjective and a common English noun. Participants were required to learn the word pairs, with the help of mediators. Mediators were either helpful, in the form of a short English definition for the adjective, or unhelpful, in the form of an equivalent Swahili definition for the adjective. Results showed a successful manipulation of mediation, as accuracy was higher for helpful mediator target retrieval (English definition supplied). However, this failed to show an interaction with the testing effect, in fact no testing effect was

found. Instead, there was indication of a negative testing effect, with the restudy only condition displaying the least deficit to memory for Swahili definition targets and the test only condition performing poorest overall.

Results were not in support of the assumptions of the mediator effectiveness hypothesis, which suggests that retrieval practice should benefit more from available mediating information in combination with retrieval practice. This study did contain a number of methodological differences from the original study in this area (Pyc & Rawson, 2010), therefore the results will be discussed in greater detail in relation to this below. It is also useful to note that the final test was completed in the same experimental session as the study phases, therefore as with experiment 2, results are also in line with evidence that testing effects can be small or reversed when the final test is completed immediately (Roediger & Karpicke, 2006b).

In experiment 4 there was also a manipulation of retrieval practice type. The original work in this area (Pyc & Rawson, 2010) included feedback in the test condition in multiple cycles of feedback and restudy, which could have contributed to the final test advantage of the test condition reported in the original study. The analysis in experiment 4 that compared test with feedback to test no feedback on final test accuracy for the two definition types, saw no evidence for an interaction between these two conditions. However, it may have been that with repeated opportunities for feedback this result would change, as repeated restudy opportunities in the form of feedback, could be particularly useful for the difficult to integrate materials utilised in experiment 4 and in Pyc and Rawson's study, by reducing working memory load (Chen et al., 2018; Van Gog & Sweller, 2015).

The results of experiment 4 were taken as evidence that mediating information alone is not enough to engender a testing effect. The results further added support to the evidence from chapter two, that no interaction between meaningful processing of the study materials and the direct effect of testing, in the absence of repeated feedback, exists.

Experiment 5 looked to explore the role of the structure of the materials as a facet

of meaningful processing. Previous work by Chan (2009) suggested that coherently organised study materials show a greater benefit of retrieval practice than less coherently organised study materials. However, in this previous work no restudy control was used. In addition, other work has previously shown evidence of the opposite trend, whereby less coherent materials benefit more from testing (de Jonge et al., 2015). New materials were devised and presented in a coherent or less coherent (random) structure but, while main effects of meaningfulness and practice task (testing effect) were found, no indication of an interaction with the testing effect was found. The results suggested that when a restudy control was added as a comparison to retrieval practice, which was absent in Chan's study, then no benefit associated with the more coherently structured text was evident. The results of experiment 5 were not consistent with the previous results (Chan, 2009; de Jonge et al., 2015), therefore a follow-up study was conducted in experiment 6 with an attempt to use a stronger manipulation of coherence.

Experiment 6 utilised more cohesive materials than experiment 5, by using short excerpts from novels that depicted a single scene, which would be more vulnerable to structural disruption. Results of the manipulation of the more cohesive study materials did not show evidence of a testing effect or the coherence manipulation. Further, there was no evidence of an interaction between these two factors. Interestingly, previous work has shown that highly cohesive materials, such as stories, do not consistently benefit from retrieval practice for the details of stories (Hostetter et al., 2019), which could be consistent with the results of de Jonge et al. (2015). Authors of this study previously suggested, that this could be due to story-like materials not being typical study materials, therefore recalling detail might not be a typical way to interact with these materials. The average accuracy results at final test in experiment 6, were the worst of all the experiments included here, despite the word stems given, suggesting there might be some merit in this perspective. In line with this, these results could be seen to fit with the notion that certain information types do not ordinarily benefit from retrieval practice. Here again therefore, it is useful to highlight that when the testing effect is absent, there is opportunity to learn something about the mechanisms of the

testing effect.

Across the three experiments in chapter three, again, there was no evidence to suggest that retrieval practice benefits in some way based on the meaningful processing of the study materials. The findings from chapters two and three were combined in a mini meta-analysis, weighted by sample size, to assess how meaningful processing of the study materials contributed to the testing effect. Combined results from experiments 1-6 revealed that the testing effect was not moderated by how meaningfully processed the study materials were, but it was moderated by the delay period.

5.2.3 Chapter Four Results

In chapter four, the focus for exploring meaningful processing changed. Based on the repeated null findings for an interaction between meaningful processing of the study materials and the testing effect, focus moved to meaningful processing achieved during the retrieval practice task. Meaningful processing was assessed for both direct retention (experiments 7 & 10) and transfer knowledge (experiments 8 & 9) effects associated with retrieval practice.

Experiment 7 assessed whether processing more of the materials previously studied during retrieval would help with retrieval of a target word. Results indicated that retrieving more of the previously studied materials benefited the testing effect in comparison to an equivalent restudy practice task. The results appeared to be due to a combination of a benefit associated with retrieving this information, alongside the detrimental impact of restudying this information. These findings suggest that increased processing during retrieval, even when that information is not necessary to final test success, increases the magnitude of the testing effect. The results of experiment 7 are consistent with a number of theories of the testing effect and will be further explored below.

Experiment 8 looked to extend the results of experiment 7 into the world of transfer knowledge, where the benefit of elaborate retrieval has shown the greatest evidence. A near transfer task was chosen, whereby the materials matched experiment 7, except

that the original retrieval practice item was a different item from the sentence. There is difficulty comparing the overall results here to experiment 7, as whilst the retrieval practice condition achieved similar results to those found in experiment 7, the restudy practice group performed at comparable levels to the retrieval practice group. Therefore, whilst no main effect of practice task or meaningful processing was found, results suggested that the increased processing task did not lead to a deficit associated with the increased processing. This could be somewhat consistent with the results in this area that have previously shown that transfer learning benefits from increased elaboration during the retrieval practice task (Endres et al., 2017; Hinze et al., 2013).

Experiment 9 explored meaningful processing by assessing whether utilising a retrieval practice task that increased comprehension might boost the effects not found in experiment 8. In experiment 9 participants retrieved a verb object instead of a noun object during retrieval practice, based on the fact that comprehension of verb objects can provide additional context to the information being learned (Barclay et al., 1974). Once again the results of experiment 9 are difficult to compare to experiment 7 results due to the higher restudy practice accuracy. The higher restudy accuracy could be a indicator that restudy practice was as effective as the retrieval practice task for the transfer test, which is somewhat consistent with previous work with associated information (Pan et al., 2016). However, as the procedure in experiment 9 for restudy participants was identical to experiment 7, we would expect the same level of accuracy. Accuracy in experiment 9 was markedly higher than experiment 7, suggesting that a boost to restudy practice is responsible. As the only differences between experiment 7 and 9 procedure for the restudy participants was running the task online, the boost is likely attributable to a number of factors associated with online studies. Possibly factors like participants studying the materials when it was convenient for them, when they had ample time and no distractions.

The absence of transfer testing effects in experiments 8 and 9 could also be due to the smaller effects associated with transfer testing effects (Adesope et al., 2017). In this way these experiments might have been further constrained by the limited dif-

ferences in the retrieval practice tasks, which could be masking the effects associated with these types of manipulations found more broadly in the literature. For example, if we compare retrieval practice tasks that differed to greater extents, for example with multiple choice questions and free recall tasks, this would reveal the differences in size of effect associated with these factors.

Experiment 10 set out to replicate the findings of experiment 7, with new materials and a modified design. Instead of elaborating based on an explanation of the plain fact in experiment 7, in experiment 10 participants instead were required to retrieve a separate word from a complementary fact. In addition, the final test was a cued fill-in-the-blank, designed to boost retrieval levels. This was to increase the likelihood of finding differences associated with the retrieval task, as opposed to processes associated with the restudy task. However, results did not replicate experiment 7. There was a testing effect demonstrated and the same numerical trend found in experiment 7, that the *what x 2* restudy trials yielded lower accuracy than the *what* only restudy trials. However, this difference was not significant and no significant interaction was seen between the two factors of interest. These results suggested that context reinstatement alone might not be enough to demonstrate differences in the testing effect and context reinstatement is likely to be beneficial when it is achieved as a function of more complete retrieval. Of course, this explanation also suggests that the reasons for the results of experiment 7 are many and not easy to untangle. It is again likely that as feedback is a feature of many key previous results in this area, feedback is able to contribute differently to finding a testing effect in its many guises.

Results from chapter four indicate that achieving more meaningful processing during retrieval might be important to the magnitude of the testing effect (experiment 7) and further might be useful in helping retrieval of transfer knowledge (experiments 8 & 9), although further work is required to reveal the nature of this relationship. With the stringent controls in place in chapter 4, the contribution of these factors is likely to be underestimated, when considering the literature more broadly. However, as the evidence in this regard was not overwhelming, a discussion of the strength of this evi-

dence will be outlined below.

5.3 Null Findings

A strength of the current work lies in the consistency of the method applied. For example, all experiments contained a retrieval practice task that was not accompanied by feedback, to enable examination of the direct effects of retrieval practice (Karpicke et al., 2014; Rowland, 2014). In addition, the studies utilised cued recall as a retrieval practice task, which typically yield the largest effects (Rowland, 2014). Furthermore, most studies employed a delay to final tests that was greater than 1 day, which also typically maximises the likelihood that testing will be more benefical than a restudy opportunity (Rowland, 2014). Yet, despite the consistency in the design elements employed, in experiments 2, 4, 6, 8 and 9, the impact of retrieval practice on overall accuracy at final test was negative or absent. There are different suggested reasons for the absence of a retrieval practice benefit in these experiments.

Firstly, results in the testing effect literature more broadly show that, when testing is immediate then the advantage of retrieval practice is reduced (Roediger & Karpicke, 2006a; Rowland, 2014). This was certainly the case in experiment 2, whereby with a delay of 5 minutes a negative testing effect was found, yet when the same study materials were tested at a delay of 3 to 5 days in experiment 3, a large testing effect was found. For ease of comparison, these results are given in figure 5.1. Therefore the delay period alone could be the contributor to the negative testing effect, although the results of experiment 1 somewhat contradict this conclusion. Experiment 1, as with experiment 2, also had a delay of 5 minutes to the final test, however in experiment 1 with the same study materials, participants completed a free recall final test and results revealed a clear testing effect. Yet, in experiment 2, with a cued recall final test results revealed a negative testing effect. Therefore, the negative testing effect, might have related not only to the short delay to the final test, but also to the inclusion of a relatively easy final test, as evidenced by the high final test accuracy rates in experiment 2 compared to those seen in experiment 1. In addition, in experiment 1

there was no significant influence of feedback to the presence of the testing effect, this pattern was also found in experiment 4. This result was rather unexpected, considering findings that have shown that testing effects can be significantly boosted by feedback alone (Rowland, 2014; van Eersel et al., 2016).

In experiment 4, the absence of a testing effect and negative testing effect, is consistent once more with the test-delay interaction, based on the fact that there was a short delay of five minutes to the final test in this experiment. However, there could also be a separate explanation for these results, based on the nature of the study materials. Experiment 4 study materials required participants to link unusual English adjectives to common English nouns, through the use of helpful (English) or unhelpful (Swahili) definitions for the unusual adjectives as mediators. Results from the complexity literature suggests that information that requires integration of various components or prior knowledge to comprehend can be evasive to testing effects (Van Gog & Sweller, 2015). As the retrieval rates were much lower than those observed in the final test in experiment 2, the absent testing effect results of experiment 4 are not likely to be solely due to a short delay to final test or a relatively easy cued recall final test. However, more work is required to assess whether the need to integrate multiple components to achieve good memory for the target words was responsible for the absent and negative testing effects seen. Interestingly, the restudy group in experiment 4 had the highest accuracy for the Swahili target words, possibly suggesting that where integration is made difficult, alternative strategies are employed such as rote memorisation (Cho & Powers, 2019). Although beyond the scope of this thesis, this explanation could be explored by varying the difficulty of integration, through difficulty associating paired items together for example, and varying the number of practice blocks to learn the information. Crucially, less work has explored these types of materials and more work is required to understand how restudy practices can be more beneficial or equally beneficial under particular circumstances.

In experiment 6, there was a lack of testing effect observed. This experiment manipulated text structure with more cohesive study materials, novel excerpts. Previous work

had demonstrated that highly cohesive materials do not benefit from testing (Hostetter et al., 2019). Experiment 6 supports this idea, possibly due to more cohesive being less memorable. However, as recent work suggests highly cohesive information benefits more from testing (Roelle & Nückles, 2019), further work is required to assess why this is the case, as no studies have manipulated structural cohesion within the same experiment.

In experiments 8 and 9, the lack of transfer effect was somewhat unexpected and could have been a function of the differences in running an experiment online. Although utilising almost identical study materials as experiment 7, performance in the restudy condition met the performance of the test condition. Therefore it is suggested that additional factors might have featured in this improved performance for the restudy group, one of which could have been increased focus during the initial session. Although it is possible, due to the nature of the tasks being transfer related, that the restudy practice generally suits this form of transfer test to a similar extent as retrieval practice. This is somewhat consistent with the idea that transfer effects tend to be smaller than retrieval practice retention tests (Pan & Rickard, 2018). Furthermore, transfer effects can be larger when the nature of the transfer task is considerably different to the retrieval practice task, for example with application and inference learning and smaller in cases where the item tested has been changed between retrieval and test as in experiments 8 and 9 (Pan & Rickard, 2018). However, results of this meta-analysis demonstrated that elaboration during retrieval was a key aspect in achieving transfer effects, although the number of studies that included elaboration were few in the sample. In experiments 8 and 9, although the retrieval item was rearranged, it was done so in the context of further elaboration during the trial. With the results of experiments 8 and 9 showing limited evidence for the contribution of elaboration to transfer effects, further work is required to assess the contribution of elaboration to specific and transfer testing effects.

Previous work has suggested that null or negative testing effects, not associated to the delay period of the final test, could be a function of design features that impair the benefits associated with retrieval practice. For example, in free recall studies, final

recall is suggested to be somewhat a feature of how well inter-item associations can be made between the items to recall. For categorically related items, when these items are presented randomly during the study phase, then presented in their categories during the practice phase, retrieval practice is impaired relative to restudy (negative testing effect) (Peterson & Mulligan, 2013). However, when the practice phase repeats the random structure presentation of the items, then a positive testing effect is found. This result is suggested to be due to both inter-item relational encoding being impaired in the restudy condition when the items are presented randomly twice and item specific encoding employed during retrieval practice outperforming restudy when inter-item processing is impaired (Mulligan & Peterson, 2015; Peterson & Mulligan, 2013). Authors have suggested that as retrieval practice enhances cue-target associations, this is why the effect is not seen when these tests are employed. Although experiment 4 employed a cued recall task, it is possible that the item-specific–relational account might predict these results due to difficulty making cue-target associations in this experiment. Further work should explore whether the difficulties associated with finding a positive testing effect with a free recall test, further extend to cued recall when materials are particularly difficult to process at the item level.

In psychological and scientific research more generally there is a notable replication crisis, whereby zeitgeist topics, that receive an avalanche of interest following impressive breakthrough results, later under scrutiny fail to replicate. In this series half of the studies reported here have not demonstrated a positive testing effect and with limited support for the highlighted theories examined, this altogether combines to suggests something of a mystery present in the testing effect literature. In this section alone I have outlined four contributing factors to finding null results, which are likely to interact with one another and currently such interactions are poorly understood. It is probably time that the field recognises that there are instances where null results are true results (Karpicke & Aue, 2015), as embracing these findings can only lead to greater understanding and a greater likelihood of a breakthrough in understanding sooner rather than later.

It has been suggested to be good practice to use meta-analytic practices to estimate the size of an effect amongst a small group of similar studies Goh et al. (2016). Therefore, a further mini meta-analysis was conducted in JASP (JASP Team, 2020) using the meta-analytic function for all testing effect comparisons in the experiments enclosed. As no reasonable evidence herein has been provided for a meaningful processing explanation of the direct effects of testing, other moderators were explored to explain the pattern of the current results. With the aim of exploring moderator effects in the mini meta-analysis, the analysis used was a Hedges random effects model (Goh et al., 2016). This computed the pooled effect size of the testing effect across all ten studies, weighted by the size of the sample in each experiment. The results revealed a pooled weighted effect size of d = 0.27 [-0.01, 0.55]. However, as the prediction interval of the current weighted effect contained zero, the resulting effect is not significant (p = .06). This average weighted effect size (although not significant) is in the small effect size range, based on standard conventions. There was an above medium level of heterogeneity in the true effects, as indicated by the I^2 value, ($50\% < I^2 = 69.85 < 75\%$) (Higgins & Thompson, 2002). The I^2 value indicates that almost 70% of total variability in the analysis is due to heterogeneity between the effects rather than to sampling error in each observed effect (Borenstein et al., 2017). This was somewhat to be expected based on the differences in effect sizes across samples. Moderator analyses explored whether the heterogeneity could be explained by previously documented phenomena in the testing effect literature. The moderators explored in the analysis were delay period to final test (same day vs. separate day) and initial accuracy (greater than 50% vs. less than 50%). Both of these factors have shown evidence of impacting the magnitude of the testing effect (Rowland, 2014). Delay period to final test did not show heterogeneity between the different levels of the moderator, p = .14, d = 0.42 [-0.14, 0.98] (same day, d = -0.03 [-0.67, 0.59] vs. separate day, d = 0.40 [0.13, 0.67]). Similarly, initial test accuracy did not show heterogeneity between levels of the moderator, p = .73, d = 0.12 [-0.53, 0.77] (accuracy > 50%, d = 0.30 [-0.08, 0.69] vs. accuracy < 50%, d = 0.21 [-0.15, 0.57]). As discussed elsewhere in this chapter, the results of the mini

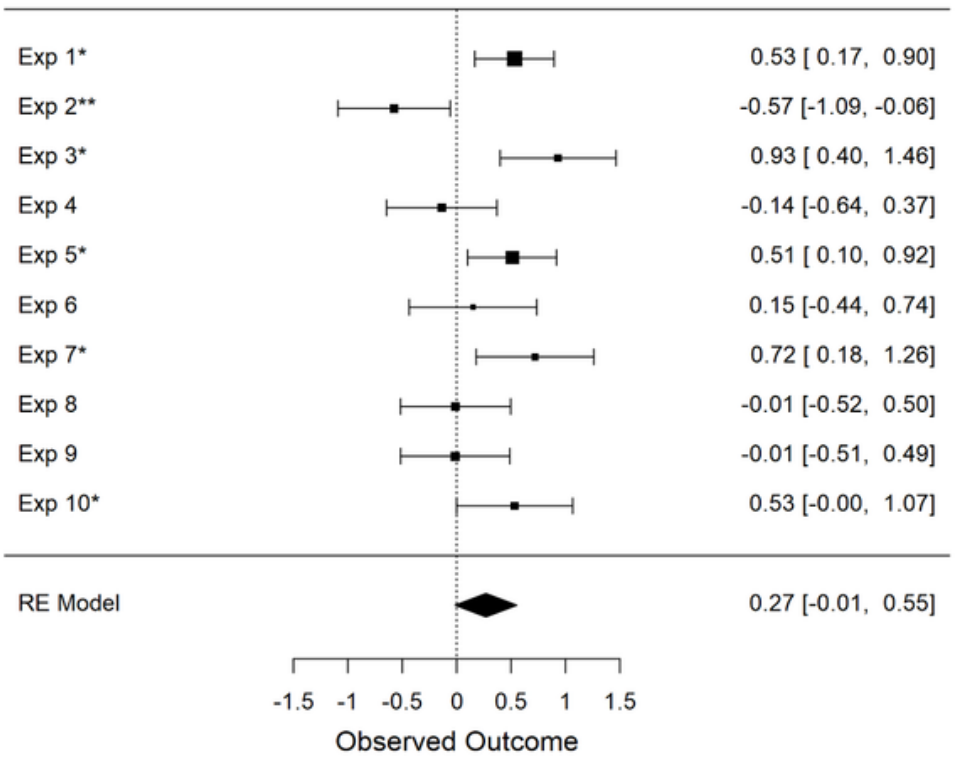

Figure 5.1. Forest plot of mini-meta analysis results, weighted by sample size for Restudy versus Test Practice. *Denotes a significant testing effect **Denotes a significant negative testing effect (restudy advantage). Exp 4 compares restudy and test without feedback.

meta-analysis further indicate that additional work is required to fully account for the pattern of results seen in the testing effect literature. The results of the meta-analysis are displayed in the forest plot in figure 5.1.

5.4 Implications for Theory

5.4.1 Elaborate Retrieval Hypothesis

Going back to the first experimental chapter, the first theory explored was the elaborate retrieval hypothesis. The three experiments from chapter two suggested that the perspective that a longer, more elaborate memory search during retrieval practice alone was not a significant contributor to the testing effect. Instead the results suggested that the original hypothesis was likely formulated from results that did not look at the impact

of retrieval practice alongside restudy, based on the compelling forgetting analysis from Carpenter (2009). These results were originally accompanied by the results of somewhat artificial practice conditions, in which the 8 item lists of word pairs were retrieved only 15 seconds after studying them (Carpenter, 2009, experiment 1). The original results could show promise for more weakly associated items being preferentially treated as a result of retrieval practice rather than a restudy task, when the conditions are right (short list, short delay, free recall test). The explanation for these results could be as a result of elaborate retrieval, or could be as a result of other phenomena such as distinctiveness properties of the items under these conditions (McDaniel & Bugg, 2008; McDaniel, DeLosh, & Merritt, 2000; Ozubko & Joordens, 2007). When the design was changed in experiment 1, to include less application limiting materials (longer list format) than those given in Carpenter, experiment 1, the testing advantage disappeared. This suggests, at best, the elaborate retrieval hypothesis does not extend to all paired associate paradigms and at worst that it is an artefact of distinctiveness properties of the materials under certain design conditions.

The fact that the delayed test in experiment 3 herein showed evidence for a lack of interaction between the meaningful processing manipulation of strong versus weak word pairs and the testing effect, is a clear suggestion that this hypothesis is not able to contribute much by way of an explanation of even a limited range of stimuli format (paired associates), as originally suggested. Furthermore, if we turn to additional experiments in this thesis, that might have been seen to explore the ERH in some looser form than the original concept, the results do not indicate any strong supporting evidence even when being creative with its application.

For example, results from experiment 4 also indicate evidence for a lack of benefit associated with the predictions of the elaborate retrieval hypothesis. Based on the fact that we would have predicted a testing effect in experiment 4, as there was the possibility in the English definition condition to make an elaborate link between the cue and target. Yet, at the final test there was no benefit associated with testing on the English definition pairs. The English definition pairs gave the suggestion of greater

elaboration through a longer search during retrieval practice. Therefore once more, it is difficult to see any evidence that elaborate processes, inherent in the studied items, indicate that a testing effect will be present.

In fact, based on the results of experiment 4, there is one piece of evidence to suggest that greater difficulty inherent in the study materials, is more likely to benefit from an additional restudy opportunity than a retrieval practice opportunity. This was the finding that for the restudy task at final test, there was no difference in final test accuracy between the low meaningful (Swahili definition) and the high meaningful (English definition) practice trials. This indicates that in some scenarios, added difficulty, or processing low in meaning materials benefits more from a restudy opportunity. These results are consistent with results from the complexity literature. Whereby, study materials that require some greater integration to comprehend, will result in an absent or negative testing effect (Van Gog & Sweller, 2015).

Strangely, the results of experiment 4 are contrary to previous criticisms of the ERH (Karpicke et al., 2014; Lehman & Karpicke, 2016), based on the finding that the elaborate restudy condition outperformed the test only condition and equalled the test with feedback condition. However, as previous work has suggested that elaborate processing is useful for learning unfamiliar concepts (Willoughby et al., 1994), this result is likely due to or resulting from an interaction with the study materials (Van Gog & Sweller, 2015), rather than being particularly insightful regarding retrieval mechanisms.

Furthermore, the results from experiment 7 could be taken as further exploration of the elaborate retrieval hypothesis, based on the fact that the ERH suggests that when more elaboration occurs during retrieval, as more effort is expended attempting to retrieve the studied information, this will provide more cues to the correct information at final test. The *what & why* trials are consistent with a more extensive memory search being conducted in retrieving the *why* element. It seems evident that this additional memory processing is useful during the retrieval practice task. However, in this case the results suggested that the benefit of both *what & why* retrieval practice, is likely due to additional processing in the retrieval practice task and a detriment associated

with additional processing in the equivalent restudy practice task. However it might be argued that in experiment 7, the opportunity for elaboration to contribute to the testing effect was somewhat constrained by the conditions of the specific cued practice and cued final tests employed. Therefore it is possible that test practices that contribute to greater variation in answers could lead to clearer differences between the retrieval practice conditions and a clearer benefit associated with retrieval practice alone. For example, perhaps a retrieval practice task that gave fewer cues and required more retrieval, versus a task that provided more cues and required less retrieval would be a sensible follow-up. However, it will also be useful to highlight the role for restudy processing or alternate processing that contributes to the effects that have been reported to date in the literature.

In addition, the results from experiments 8 and 9 suggest once more that the presence of elaboration during retrieval is not reason alone to find a testing effect, even when retrieval rates are high. For example, no testing effect was seen in either experiment, despite the fact that elaborate retrieval was present in retrieval practice and additional relational processing was included (experiment 9). In addition, while comprehension is likely to be boosted in the retrieval task of experiment 9 based on the increased relational processing associated with verb retrieval, still no testing effect was found. Elaborate retrieval relies on the idea of increased semantic processing during retrieval practice being useful for subsequent retrieval (Carpenter, 2009, 2011). Therefore it is surprising that no testing effect was found in particular in experiment 9. Although, had there been a testing effect found in experiment 9 it might have equally been explained by increased difficulty associated with this task, as during the practice tasks the verb in experiment 9 took longer to retrieve than the nouns in experiment 8.

Therefore the results of the experiments included herein do not provide any clear evidence to suggest that elaborate retrieval is responsible for the testing effect, instead it is likely that the results associated with the original formulation of this hypothesis are due to special design conditions. Further work might look to assess whether distinctiveness during free recall is responsible for the findings associated with Carpenter's

2009 work and whether such conditions might be utilised to make use of any advantage that might be applied to learning weakly associated items.

5.4.2 Mediator Effectiveness Hypothesis

In chapter three the mediator effectiveness hypothesis was also explored. The mediator effectiveness hypothesis suggests that available mediating information during the retrieval practice task assists retention and subsequent retrieval at the final test. In experiment 4 this was explored by adding in a definition for an unusual English cue word that was being paired with a common noun object. Participants were tasked with making a link between the cue and target during the study phases, to be able to successfully retrieve the target from the cue in the retrieval practice task. The meaningful processing manipulation was successful, with accuracy for the English definition word pairs being greater than accuracy for the Swahili definition word pairs. Experiment 4 utilised two separate versions of a retrieval practice task, one which included feedback and one which did not. Neither of these two conditions suggested that there was a retrieval practice benefit associated with the high mediator (English definition) word pairs, suggesting that the mediator account is not on its own sufficient to understanding the testing effect.

The format of the materials of experiment 4 was very similar to the format of the study items in Pyc and Rawson (2010), from which the mediator effectiveness hypothesis has come. Yet, some changes in the design are observed between experiment 4 and Pyc and Rawson's study, in which multiple retrieval-feedback cycles were utilised during the practice phase. The differences that might be associated with these design differences will be discussed below.

In Pyc and Rawson's study all participants experienced the word pairs with a mediator, self-generated, four times, an initial study trial followed by three practice phases. In each retrieval practice phase, participants attempted to recall the target from the cue. Following each retrieval attempt, participants were given the word pair intact and asked to provided the previously generated mediator for the pair. In the restudy practice

tasks participants were asked to generate the mediator, when restudying the word pair intact. Based on the original results, some of the claims for the mediator effectiveness hypothesis lack the strength in evidence for a benefit associated with retrieval practice.

There were three conditions at the final test. There was a large discrepancy between the restudy and retrieval conditions on final test accuracy (retrieval practice advantage) when only the cue was provided and the target was asked for at the final test. This discrepancy was markedly reduced when the mediator is also provided. The discrepancy is somewhat between these two when the task is to retrieve both the original mediator and the target. The authors suggest that mediators are being better utilised when teamed with retrieval practice than restudy. The evidence given in the original paper supports this specific claim, due to all final test conditions resulting in better performance in the retrieval practice condition. What this result does not suggest is that retrieval practice enables better spontaneous activation of mediating information. It merely suggests that when you double the number of practice opportunities, you see a memorial benefit for all information studied. Interestingly, when mediators are not a part of the final test, then the difference between retrieval practice and restudy practice is largest. This possibly provides evidence for transfer appropriate processing or desirable difficulties associated with retrieval practice.

However, the results of experiment 4 suggest that when mediating information is part of the study materials only, then there is no advantage to retrieval practice without explicit instruction to retrieve the mediating information. So the original claim might need to be revisited. It is not clear whether Pyc and Rawson (2010) are suggesting that retrieval boosts the utility of mediating information or that mediating information boosts the utility of retrieval practice. Either way, without comparing the original results to a condition in which no mediation was present to help participants to remember, it is difficult to go beyond the specific conditions in which the results were found. Experiment 4 provides that opportunity. By not providing the mediating information during retrieval practice, results showed that mediating information is more beneficial to the likelihood of retrieval than less mediating information. Yet, results did not indicate that mediating

information is beneficial to the testing effect.

5.4.3 Constructed Retrieval

Constructed retrieval was not explicitly explored, but has been used as an explanation for retrieval practice results that have benefited transfer knowledge (Hinze et al., 2013). Experiment 8 tested the idea that in relation to transfer knowledge, elaboration that helped to construct knowledge helped to preserve accuracy for the target information. For experiment 8 this would be the *why* element requiring active construction of knowledge relevant to retrieving the target information contained in the *what* element. The *what & why* trials did not lead to superior performance than the *what* only trials. Therefore, the benefit of constructed retrieval practice might be limited in the case that the final test contains very limited transfer. This result is consistent with the results of experiment 9 also. However, based on the findings of experiment 7 and 10, constructing retrieval might be a useful component of when testing will be most beneficial. Yet, the results of experiments 7 and 10 that examined retention through testing and not transfer, constructed retrieval would be an adequate explanation for when testing was more beneficial (Hinze et al., 2013). As increased opportunity to construct knowledge in relation to the target information in experiment 7 led to a boost in the testing effect, yet when this opportunity was reduced in experiment 10 no increase in the testing effect was seen. However, the results of experiments 7 & 10 could also be explained by elements of the elaborate retrieval hypothesis and the episodic context account equally. Future work will need to address whether these factors can be disassociated.

5.4.4 Bifurcated Distribution

From the meta-analysis results conducted in this chapter, a direct assessment of the impact of initial test accuracy to the magnitude of the testing effect has been made across the 10 experiments given here, based on the moderator analysis. The results of the moderator analysis of initial test accuracy during retrieval contributing to the testing effect, suggests that there is not evidence in the experiments enclosed that initial test accuracy alone is a significant predictor of the testing effect. The results of the mod-

erator analysis for whether delay moderates the testing effect, suggested that across all 10 experiments this was not a significant predictor of the testing effect, which is another feature suggested to be important in the bifurcated distribution account (Kornell et al., 2011). However, the results of the moderator analysis in the mini meta-analysis conducted on the experiments from chapters two and three suggested that delay was a significant indicator of the testing effect. Taken together and in light of the factors already discussed, these results suggest the bifurcated distribution alone does not seem to adequately account for the findings outlined in the experiments enclosed, but is still able to describe results in some cases.

5.4.5 Episodic Context Account

One theory of the testing effect suggests that the benefit associated with the practice test is influenced by the extent to which the original study episode can be reinstated (Karpicke et al., 2014). Reinstating the context of the original episode during retrieval allows the cues to be updated and strengthened for subsequent retrieval. In line with this, the *what & why* trials in experiment 7, encouraged more reinstatement of the original study episode by requiring that participants attempt to retrieve more of the originally studied item than required in the *what* only retrieval trials. Therefore one factor contributing to the results reported could be that initial reinstatement of the earlier study episode contributed to better long-term memory benefit. This result was explored in more detail in experiment 10, whereby increased context reinstatement was achieved in the *what x 2* trials, by re-presenting a large portion of the original study item in combination with an additional retrieval opportunity. In experiment 10, no additional benefit to the testing effect was associated with *what x 2* retrieval, which is not consistent with the episodic context account. However, it may be that the reinstatement needs to be achieved more completely through retrieval to see any benefits, as was seen in experiment 7. As the explanation for these findings can equally be explained by constructed retrieval or elaborate processing, future work should examine to what extent these elements are dissociable and are in fact contributors. A recent study found that

reinstating aspects of the context during retrieval context, for example, the position the item was presented in, led to better comprehension but not retention (van den Broek, Takashima, Segers, & Verhoeven, 2018). Therefore further research should assess whether context reinstatement is beneficial to retention, as could be indicated by the results of experiment 7.

In addition, in experiment 4 findings were not consistent with previous results and claims of the episodic context account, namely findings that elaborate restudy is inferior to retrieval practice but should be equal if there is any merit in the elaborate retrieval account (Lehman & Karpicke, 2016). As the elaborate restudy condition outperformed the test only condition and performed in line with the test with feedback condition, there are at least some instances where elaboration during study is more useful than testing. It is likely that these results indicate that deeper processing can be useful for learning novel information (Willoughby et al., 1994), rather than explicit support for the elaborate retrieval hypothesis.

5.5 Implications for Practice

Always at the forefront of my considerations during the exploration of the contribution of meaningful processing to retrieval practice effects was how these results can translate into something meaningful for educational practice. The consensus in the literature is that testing is a useful tool to employ for boosting long-term memory for educational materials we wish to learn. The results given here do not contest this notion, as many studies have demonstrated good testing effects. Across the ten experiments given here, the sample weighted average of the 10 effect sizes in the mini-meta analysis gives the overall effect size of $H = .27$, for retention via cued recall across a range of materials.

Many factors that contribute to the testing effect have been discussed. One such aspect relevant to implications for practice relates to the degree of integration required to comprehend the study materials, as was explored in chapter three. This aspect of comprehension relates to work that has been done on how prior knowledge influences

the testing effect. In educational research, prior knowledge is the foundation for comprehension and building more complex associative representations (Elleman & Compton, 2017). It is this that is thought to differ between poor and good comprehenders. In an early study, Schneider et al. (1989) found that low aptitude students could recall information just as well as high aptitude pupils when the subject area was an area of expertise, expertise was related to recall performance, regardless of aptitude. Elleman and Compton (2017) reviewed factors important to reading comprehension in children, and suggested that the relevance of student's prior knowledge has been consistently shown in the literature, but has been somewhat overlooked in recent times. These results are consistent with findings relating to text cohesion (O'Reilly & McNamara, 2007), whereby authors suggest that addressing the importance of prior knowledge would be a useful focus for future work. The importance of individual differences in knowledge has been speculated for some time (Morris et al., 1977; Moscovitch & Craik, 1976) in relation to memory performance and is something that is suggested for further enquiry. Therefore it is suggested here that work examining how ability to integrate study information should be further explored, as this is likely to have a large impact on reaching individuals who are most disadvantaged in an educational environment.

Further implications for practice come from the findings that differences in meaningful processing of the study materials do not directly impact the magnitude of the testing effect. This is important because study information is often made up of a mix of information that varies in how meaningful it is. Results from chapters three and four suggest that test practice is mutually beneficial to the learning of items that are more meaningfully processed and less meaningfully processed when compared to a restudy control, at least when the items are studied as part of the same study materials. Although items that are more meaningfully processed benefit more from both types of practice, based on high meaningful items showing higher final test accuracy. High and low meaningful items were mostly (except for experiment 5) combined within-subjects herein as this can typically exaggerate effects in memory. For example, generation effects rely on within-subject manipulations (Ozubko & MacLeod, 2010). However, ev-

idence across experiments suggests that some features of the study materials might influence the testing effect. In particular, results of chapter three saw different materials utilised and differences in whether the testing effect was present (experiment 5) or absent (experiment 4 & 6). As previous research has indicated that structural properties of the information can alter the testing effect (Chan, 2009; de Jonge et al., 2015; Hostetter et al., 2019; Roelle & Nückles, 2019; Rowland, 2014; Van Gog & Sweller, 2015), future work should look to see whether between-subject comparisons could be the reason these differences have previously been found. This will further our understanding of when testing is beneficial and when restudy is equally beneficial.

Another implication for practice is that no positive testing effect was observed in half of the experiments detailed here. This suggests, that although retrieval practice is particularly beneficial when high accuracy or feedback is a feature of the retrieval practice task, some circumstances demonstrate an equivalent benefit to learning for restudy and retrieval practice. As the boundary conditions for this effect are not comprehensively understood, properties of restudy tasks should be evaluated for a better understanding of the testing effect. Furthermore, restudy may not be as useless a learning strategy as implied by the field, and could even be preferable in some cases (experiment 4).

5.6 Future Work

One suggestion for future work should be for a greater exploration of the role of integration. The current results encouragingly suggest that meaningful processing in the study materials is not likely to impact on the magnitude of the testing effect and is generally uniformly benefited from a retrieval opportunity in comparison to a restudy opportunity. However, crucially the results from chapter three should be qualified in relation to the fact that the nature of the materials contained information that ranged in how easy to integrate it was. Consistent with previous work results here suggested that easy to integrate materials might not benefit from testing (experiment 6, de Jonge et al., 2015) in the same way that less easy to integrate information does (experiment 5, de Jonge

et al., 2015). Yet materials arguably difficult to integrate, in experiment 4, showed no benefit of testing. These results could be somewhat consistent with findings that individuals learn within a zone of proximal development and as such individual testing schedules have been shown to be useful (Metcalfe & Kornell, 2003).

Worth consideration for future work from the results explored within this thesis, is the role of the restudy task. The restudy task is primarily used in testing effect studies due to its application to educational practice. The restudy task acts as a baseline measure for comparison to the retrieval practice task, typically utilised as a passive re-exposure task with matched time to the retrieval practice task. The specific benefit of retrieval practice for long-term memory could be further understood when considering restudy task performance. Chapter two results highlighted that properties of retrieval practice alone could not reliably predict the pattern of results at the final test when compared to a restudy task. For example, initial test properties, like response times and accuracy levels were not a good indicator of the benefit of retrieval practice over restudy practice for the testing effect (experiment 1). In addition, in chapter three the results of experiments 5 and 6 suggested that items that differ in meaningful processing may appear to relate to differences in processing during retrieval practice, but only when not accompanied by a comparable restudy control (Chan, 2009). But this evidence taken together suggests that meaningful processing is likely represented in the same way by restudy and retrieval practice, with items that are more meaningfully processed being subjected to forgetting in the same way that items that are less meaningfully processed are.

In relation to the elaboration that occurs during the retrieval practice task, here it is suggested that this is likely to contribute to the testing effect. However, this could be for a number of reasons; due to the complementary processes on an equivalent restudy task, due to whether the final test contains a measure of transfer learning, or possibly due to the extent to which the previous study episode has been reinstated, elaborated on or constructed in one's own memory. Future work should look to explore these explanations, as well as when feedback is particularly useful at boosting these

effects.

5.7 Conclusion

Across ten experiments the impact of meaningful processing in both the study materials and the retrieval practice task was explored. Results revealed that changes in meaningful processing in the study materials are not likely to influence the testing effect, however further work should explore aspects of cohesion or ability to integrate the materials as a contributor to the testing effect. In addition, the present results found that more meaningful processing during the retrieval practice task can contribute to the testing effect (experiment 7). This could be due equally to a benefit associated with meaningful processing in the retrieval practice task and a deficit associated with an equivalent study task. There could be a case for greater context reinstatement from the follow-up why questions showing the benefit, however further work needs to isolate the effects associated with semantic and non-semantic (contextual) processing. Where feedback was included, this did not appear to significantly alter the magnitude of the testing effect, however the present results and previous work indicate this to be a key area for future work. The present results indicate that more attention should be paid to the restudy control task and what this might reveal about retrieval practice. In particular, this could be achieved by matching the restudy task as closely as possible to the retrieval practice task. Further investigation should in this regard should avoid assigning properties of greater practice in general to properties of retrieval practice specifically.

List of references

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701.

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131–139.

Akan, M., Stanley, S. E., & Benjamin, A. S. (2018). Testing enhances memory for context. *Journal of Memory and Language*, *103*, 19–27.

Allen, G. A., Mahler, W. A., & Estes, W. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *8*(4), 463–470.

Anderson, J. R., & Reder, L. M. (1979). An elaborative processing explanation of depth of processing. *L.; S. Cermak and FIM Craik, Eds., Levels of Processing in Human Memory (Erlbam, 1979)*, 385–404.

Anderson, M. C., & McCulloch, K. C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(3), 608.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, *52*(1), 388–407.

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, *13*(4), 471–481.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, *128*(4), 612.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive psychology*, *61*(3), 228–247.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition*, *35*(2), 201–210.

Bjork, E. L., Bjork, R. A., et al. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Bjork, R. A. (1999). *F 5 assessing our own competence: Heuristics and illusions.*

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*, 417–444.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research synthesis methods*, *8*(1), 5–18.

Bouwmeester, S., & Verkoeijen, P. P. (2011). Why do some children benefit more from testing than others? gist trace processing to explain the testing effect. *Journal of Memory and Language*, *65*(1), 32–41.

Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 165–174.

Bregman, A. S., & Wiener, J. R. (1970). Effects of test trials in paired-associate and free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *9*(6), 689–698.

Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory*, *18*(4), 385–393.

Buck, S., Ritter, G. W., Jensen, N. C., & Rose, C. P. (2010). Teachers say the most interesting things—an alternative view of testing. *Phi Delta Kappan*, *91*(6), 50–

54.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514–527.

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30–41.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & cognition*, *34*(2), 268–276.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448.

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141.

Chan, J. C. (2009). When does retrieval induce forgetting and when does it induce facilitation? implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*(2), 153–170.

Chan, J. C. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*(1), 49–57.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Undesirable difficulty effects in the learning of high-element interactivity materials. *Frontiers in psychology*, *9*, 1483.

Cho, K. W., Neely, J. H., Brennan, M. K., Vitrano, D., & Crocco, S. (2017). Does testing increase spontaneous mediation in learning semantically related paired associates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1768.

Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory and Cognition*, *8*(2), 166–177.

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition*, *2*(2), 95–100.

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505.

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition*, *40*(4), 528–539.

Coppens, L. C., Verkoeijen, P. P., Bouwmeester, S., & Rikers, R. M. (2016). The testing effect for mediator final test cues and related final test cues in online and laboratory experiments. *BMC psychology*, *4*(1), 25.

Craik, F. I. (1970). The fate of primary memory items in free recall. *Journal of verbal learning and verbal behavior*, *9*(2), 143–148.

Craik, F. I. (2002). Levels of processing: Past, present... and future? *Memory*, *10*(5-6), 305–318.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*, *104*(3), 268.

Crouse, J. H. (1967). Free learning as a function of meaningfulness and encoding cues. *Psychonomic Science*, *7*(10), 361–362.

de Jonge, M., Tabbers, H. K., & Rikers, R. M. (2015). The effect of testing on the retention of coherent and incoherent text material. *Educational Psychology Review*, *27*(2), 305–315.

deWinstanley, P. A., & Bjork, R. A. (2002). Successful lecturing: Presenting information in ways that engage effective processing. *New directions for teaching and learning*, *2002*(89), 19–31.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228–232.

Ebbinghaus, H. (1913). Memory: A contribution to experimental psychology (ar ruger & ce bussenius, trans.). new york: Teachers college, columbia university.(original work published 1885). *Memory & Cognition*, *12*, 105–111.

Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*(1), 215–228.

Einstein, G. O., McDaniel, M. A., Owen, P. D., & Cote, N. C. (1990). Encoding and recall of texts: The importance of material appropriate processing. *Journal of Memory and Language*, *29*(5), 566–581.

Elleman, A. M., & Compton, D. L. (2017). Beyond comprehension strategy instruction: What's next? *Language, Speech, and Hearing Services in Schools*, *48*(2), 84–91.

Endres, T., Carpenter, S., Martin, A., & Renkl, A. (2017). Enhancing learning by retrieval: Enriching free recall with elaborative prompting. *Learning and Instruction*, *49*, 13–20.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*(41), 1149–1160.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas,

and thematic role concepts. *Journal of Memory and Language*, *44*(4), 516–547.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2-3), 285–307.

Gates, A. I. (1922). *Recitation as a factor in memorizing* (No. 40). Science Press.

Gentner, D., & France, I. M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In *Lexical ambiguity resolution* (pp. 343–382). Elsevier.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (2000, 01). Human simulations of vocabulary learning. *Cognition*, *73*, 135-76. doi: 10.1016/S0010-0277(99)00036-0

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology*, *66*(3), 325–331.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549.

Goodmon, L. B., & Anderson, M. C. (2011). Semantic integration as a boundary condition on inhibitory processes in episodic retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 416.

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*, 2412.

Hadidi, Y., & Nazerfar, R. (2014). Comments on the system of lexical cohesion in a sample of english fiction". *2014*.

Hakim, R. A. (2016). Lexical cohesion in fiction stories with reference to the frog prince and the bully.

Hausman, H., & Rhodes, M. G. (2018). Retrieval activates related words more than presentation. *Memory*, *26*(9), 1265–1280.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis.

Statistics in medicine, *21*(11), 1539–1558.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290–304.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151–164.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562–567.

Hostetter, A. B., Penix, E. A., Norman, M. Z., Batsell Jr, W. R., & Carr, T. H. (2019). The role of retrieval practice in memory and analogical problem-solving. *Quarterly Journal of Experimental Psychology*, *72*(4), 858–871.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Hulme, C., Maughan, S., & Brown, G. D. A. (1991, Dec 01). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, *30*(6), 685. Retrieved from https://www.proquest.com/scholarly-journals/memory-familiar-unfamiliar-words-evidence-long/docview/1297337597/se-2?accountid=14711 (Last updated - 2013-02-23)

Jarosz, A. F., & Wiley, J. (2014). What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

JASP Team. (2020). *JASP (Version 0.9.2)[Computer software].* Retrieved from https://jasp-stats.org/

Johnson-Laird, P., Gibbs, G., & De Mowbray, J. (1978). Meaning, amount of processing, and memory for words. *Memory & Cognition*, *6*(4), 372–375.

Kang, S. H., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528–558.

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. *Grantee Submission*.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of learning and motivation* (Vol. 61, pp. 237–284). Elsevier.

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of experimental psychology: learning, memory, and cognition*, *33*(4), 704.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*(1), 17–29.

Kintsch, W. (1974). The representation of meaning in memory.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989.

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013). The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education*, *18*(3), 409–425.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-

term retention relative to repeated study: a randomised controlled trial. *Medical education*, *43*(12), 1174–1181.

Lechuga, M. T., Ortega-Tudela, J. M., & Gómez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. *Learning and Instruction*, *40*, 61–68.

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1573.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American psychologist*, *63*(8), 760.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review*, *15*(2), 237–255.

McDaniel, M. A., DeLosh, E. L., & Merritt, P. S. (2000). Order information and retrieval distinctiveness: Recall of common versus bizarre material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 1045.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*(4), 516–522.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371.

McDaniel, M. A., & Pressley, M. (1989). Keyword and context instruction of new vocabulary meanings: Effects on text comprehension and memory. *Journal of Educational Psychology*, *81*(2), 204.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time

to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*(4), 530.

Morey, R. D., et al. (2008). Confidence intervals from normalized data: A correction to cousineau (2005). *reason*, *4*(2), 61–64.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of verbal learning and verbal behavior*, *16*(5), 519–533.

Moscovitch, M., & Craik, F. I. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of verbal learning and verbal Behavior*, *15*(4), 447–458.

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 859.

Mulligan, N. W., Susser, J. A., & Smith, S. A. (2016). The testing effect is moderated by experimental design. *Journal of Memory and Language*, *90*, 49–65.

Nathoo, F. S., Kilshaw, R. E., & Masson, M. E. (2018). A better (bayesian) interval estimate for within-subject designs. *Journal of Mathematical Psychology*, *86*, 1–9.

Nelson, D., McEvoy, C., & Schreiber, T. (1998). *The university of south florida word association, rhyme, and word fragment norms. 1998 http://www. usf. edu.* Free-Association.

Ozubko, J. D., & Joordens, S. (2007). The mixed truth about frequency effects on free recall: Effects of study list composition. *Psychonomic Bulletin & Review*, *14*(5), 871–876.

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1543.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures

of high school students' science achievement. *American educational research journal*, *44*(1), 161–196.

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied*, *23*(3), 278.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological bulletin*, *144*(7), 710.

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*, *44*(1), 24–36.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195–203.

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1287.

Psychology Software Tools, Inc, Pittsburgh, PA. (2016). *E-Prime 3.0.* Retrieved from https://www.pstnet.com/

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*(8), 1917–1927.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335–335.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619–633.

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition:

The impetus from the levels-of-processing framework. *Memory*, *10*(5-6), 319–332.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, *1*(3), 181–210.

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

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: the complexity of the tasks matters. *Learning and Instruction*, *49*, 142–156.

Roelle, J., & Nückles, M. (2019). Generative learning versus retrieval practice in learning from text: The cohesion and elaboration of the text matters. *Journal of Educational Psychology*, *111*(8), 1341.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432.

Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed-versus pure-list designs. *Memory & cognition*, *42*(6), 912–921.

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied*, *23*(3), 293.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, *11*(6), 641–650.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 26.

Schneider, W., Körkel, J., & Weinert, F. E. (1989). Domain-specific knowledge and memory performance: A comparison of high-and low-aptitude children. *Journal of educational psychology*, *81*(3), 306.

Schwoebel, J., Depperman, A. K., & Scott, J. L. (2018). Distinct episodic contexts enhance retrieval-based learning. *Memory*, *26*(9), 1291–1296.

Shimmerlik, S. M. (1978). Organization theory and memory for prose: A review of the literature. *Review of Educational Research*, *48*(1), 103–120.

Smith, M. A., Blunt, J. R., Whiffen, J. W., & Karpicke, J. D. (2016). Does providing prompts during retrieval practice improve learning? *Applied Cognitive Psychology*, *30*(4), 544–553.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition*, *6*(4), 342–353.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review*, *8*(2), 203–220.

Soraci, S. A., Franks, J. J., Bransford, J. D., Chechile, R. A., Belli, R. F., Carr, M., & Carlin, M. (1994). Incongruous item generation effects: A multiple-cue perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 67.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of experimental psychology*, *70*(1), 122.

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., ... others (2020). A tutorial on conducting and interpreting a bayesian anova in jasp. *LAnnee psychologique*, *120*(1), 73–96.

van den Broek, G. S., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, *68*(2), 546–585.

van Eersel, G. G., Verkoeijen, P. P., Povilenaite, M., & Rikers, R. (2016). The testing effect and far transfer: The role of exposure to key information. *Frontiers in*

Psychology, *7*, 1977.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247–264.

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580.

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1036.

Willoughby, T., Wood, E., & Khan, M. (1994). Isolating variables that impact on or detract from the effectiveness of elaboration strategies. *Journal of Educational Psychology*, *86*(2), 279.

Wilson, M. (1988). Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, *20*(1), 6–10.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & cognition*, *38*(8), 995–1008.

Appendix A

Chapter Two Materials

A.1 Word pairs for Exps 1, 2 & 3

Practice Items	
Cue	Target (strong)
Icing	Sugar
Omelette	Eggs

Main Item List				
Cue	Strong target	Strength of association	Weak target	Strength of association
Barrier	Wall	0.329	Bridge	0.014
Beach	Sand	0.394	Shore	0.012
Beast	Animal	0.4	Savage	0.012
Brake	Stop	0.412	Pedal	0.014
Bucket	Water	0.373	Kick	0.014
Calcium	Milk	0.413	Tablets	0.013
Calendar	Date	0.305	Schedule	0.013
Dentist	Teeth	0.459	Cavity	0.02
Chime	Bell	0.36	Ding	0.013
Chisel	Hammer	0.328	Nail	0.01
Chore	Work	0.408	Housework	0.011
Cloak	Dagger	0.441	Vampire	0.013
Crane	Bird	0.318	Steel	0.014
Crease	Fold	0.308	Skirt	0.014
Cycle	Bike	0.318	Wheel	0.014
Designer	Clothes	0.427	Shoes	0.013
Diner	Food	0.318	Waiter	0.013
Dresser	Drawer	0.43	Wardrobe	0.013
Entry	Exit	0.387	Rear	0.013
Ethics	Morals	0.331	World	0.014
Fraud	Fake	0.319	Scandal	0.014
Gravy	Potato	0.313	Roast	0.016
Grove	Orange	0.436	Peach	0.012
Headache	Pain	0.361	Fever	0.014
Lobby	Hotel	0.345	Smoke	0.014
Margin	Paper	0.322	Victory	0.013
Melody	Song	0.414	Harmony	0.093
Monument	Statue	0.364	Tower	0.014
Mosquito	Bite	0.362	Sting	0.014
Needle	Thread	0.424	Prick	0.012
Pajamas	Sleep	0.359	Gown	0.014
Patrol	Police	0.367	Watchdog	0.013
Pottery	Clay	0.384	Plant	0.013
Pouch	Kangaroo	0.303	Wallet	0.01
Prank	Joke	0.385	Fool	0.01

Word pairs for Exps 1, 2 & 3 continued

Priest	Church	0.382	Bible	0.014
Rainbow	Colour	0.357	Cloud	0.014
Remedy	Cure	0.429	Relief	0.014
Rhythm	Beat	0.354	Tempo	0.01
Rider	Horse	0.372	Saddle	0.014
Ruler	Measure	0.415	Metre	0.015
Shelf	Book	0.368	Stuff	0.013
Shrine	Temple	0.208	Idol	0.013
Smudge	Smear	0.434	Mess	0.014
Suburb	City	0.265	Ghetto	0.013
Sweep	Broom	0.406	Carpet	0.022
Sword	Knife	0.301	Spear	0.013
Topic	Subject	0.385	Headline	0.014
Vessel	Ship	0.347	Vein	0.014
Violin	Music	0.372	Orchestra	0.013
Wicker	Basket	0.304	Rattan	0.027
Alcohol	Beer	0.24	Vodka	0.014
Apron	Cook	0.291	Smock	0.014
Chart	Graph	0.275	Statistics	0.013
Chemist	Scientist	0.262	Physicist	0.016
Ditch	Hole	0.276	Weed	0.02
Dragon	Fire	0.275	Scales	0.014
Festival	Party	0.277	Rides	0.014
Flannel	Shirt	0.244	Cloth	0.016
Gauze	Bandage	0.203	Wrap	0.014
Gravel	Rocks	0.25	Quarry	0.013
Necklace	Gold	0.217	Broach	0.014
Outlet	Plug	0.217	Cord	0.013
Plaza	Mall	0.23	Centre	0.014
Trumpet	Horn	0.245	Tuba	0.014
Utensil	Fork	0.328	Silverware	0.017
Convict	Jail	0.194	Villain	0.014
Pickle	Cucumber	0.164	Prune	0.014
Stairway	Heaven	0.189	Railing	0.014
Surgery	Operation	0.158	Scar	0.014
Coyote	Wolf	0.237	Moon	0.014
Radish	Vegetable	0.26	Sprout	0.014
Explorer	Adventure	0.196	Scout	0.014
Fighter	Boxer	0.209	Fist	0.014
Diary	Secrets	0.237	Memoirs	0.013
Meadow	Field	0.299	Valley	0.014
Symbol	Sign	0.186	Signal	0.014
Shave	Razor	0.151	Lather	0.014
Custard	Pudding	0.263	Doughnut	0.013
Mustard	Ketchup	0.584	Spice	0.019

Appendix B

Chapter Three Materials

B.1 Word pairs for Exp 4

Practice Items				
Item No	**Adjective**	**Noun**	**English definition**	**Swahili definition**
1	**Barbarous**	**Litter**	rough and uncivilized	mbaya na isiyostahili
2	**Unsullied**	**Saucer**	pure, unspoiled, not linked with unpleasantness	safi, haijatilishwa, haihusiani na unpleasantness
Main Items				
Item No	**Adjective**	**Noun**	**English definition**	**Swahili definition**
3	**Propitious**	**Shrub**	gracious or favourably inclined	neema au mzuri
4	**Hoary**	**Turtle**	old and familiar	zamani na ya kawaida
5	**Officious**	**Arrow**	interfering, eager to tell people what to do	kuingilia kati, nia ya kuwaambia watu nini cha kufanya
6	**Impious**	**Tulip**	showing a lack of respect for religious things	kuonyesha ukosefu wa heshima kwa mambo ya dini
7	**Assiduous**	**Napkin**	hard working, thorough	kufanya kazi ngumu, vizuri
8	**Querulous**	**Mattress**	likes to moan and complain	anapenda kusuhi na kulalamika
9	**Blithe**	**Buckle**	casual, carefree, thoughtless	kawaida, wasiwasi, wasiwasi
10	**Sepulchral**	**Almond**	serious, sad, frightening	mbaya, kusikitisha, kutisha

Word pairs for Exp 4 continued

11	**Capricious**	**Pigeon**	unpredictable, impulsive	haitabiriki, hasira
12	**Obdurate**	**Muffin**	unreasonable, stubborn	wasio na busara, mkaidi
13	**Commodious**	**Monkey**	huge, spacious, large capacity	kubwa, wasaa, uwezo mkubwa
14	**Doughty**	**Ornament**	brave, determined, a fighter	jasiri, kuamua, mpiganaji
15	**Inveterate**	**Squirrel**	something habitual, not likely to change	kitu cha kawaida, sio uwezekano wa kubadili
16	**Sagacious**	**Patio**	intelligence, wisdom	akili, hekima
17	**Nefarious**	**Tissue**	sinful, immoral	wenye dhambi, uovu
18	**Cogent**	**Canal**	convincing, makes good sense	kushawishi, hufanya akili nzuri
19	**Virulent**	**Barrel**	extremely bitter and hostile	uchungu sana na chuki
20	**Garrulous**	**Shield**	talkative, talks about unimportant things	kuongea, huzungumzia mambo yasiyo muhimu
21	**Ignominious**	**Scarf**	embarrassing, failing miserably	aibu, kushindwa sana
22	**Inalienable**	**Cinema**	rights that cannot be changed or taken away	haki ambazo haziwezi kubadilishwa au kuondolewa
23	**Incipient**	**Guitar**	just starting to happen	kuanza tu kutokea

Word pairs for Exp 4 continued

24	**Incorrigible**	**Parcel**	has faults that will never change	ina makosa ambayo hayawezi kubadilika
25	**Indolent**	**Mouse**	lazy, does not like to work	wavivu, hapendi kufanya kazi
26	**Ineffable**	**Statue**	so great or extreme it cannot be described in words	hivyo kubwa au uliokithiri hauwezi kuelezewa kwa maneno
27	**Inimical**	**Blouse**	makes it difficult for something else to exist or do well	inafanya kuwa vigumu kwa kitu kingine kuwepo au kufanya vizuri
28	**Inimitable**	**Wallet**	has qualities you admire	ina sifa unazozipenda
29	**Limpid**	**Fiddle**	clear and transparent	wazi na uwazi
30	**Abstruse**	**Orchard**	difficult to understand	vigumu kuelewa
31	**Benighted**	**Coffin**	ignorant, lacking culture	wasiojua, kukosa utamaduni
32	**Nebulous**	**Arena**	vague, not easy to describe	wazi, si rahisi kuelezea
33	**Ostentatious**	**Lemonade**	expensive, impressive	ghali, ya kushangaza

Word pairs for Exp 4 continued

34	**Voracious**	**Crystal**	hungry, has a large appetite for something	njaa, ina hamu kubwa ya kitu
35	**Pernicious**	**Anchor**	very harmful	madhara sana
36	**Parochial**	**Shovel**	too focused on personal or unimportant things	pia ililenga vitu vya kibinafsi au visivyofaa
37	**Auspicious**	**Donkey**	likely to succeed	uwezekano wa kufanikiwa
38	**Pugnacious**	**Custard**	always ready to quarrel or fight	daima tayari kupigana au kupigana
39	**Punctilious**	**Rocket**	very careful to behave correctly	makini sana kufanya vizuri
40	**Recalcitrant**	**Award**	unwilling to obey orders, difficult to work with	hawataki kuitii amri, vigumu kufanya kazi na
41	**Titular**	**Camel**	has a name that sounds important but is not really important	ina jina linaloonekana kuwa muhimu lakini si muhimu sana
42	**Obsequious**	**Hedge**	obedient, eager to please	watiifu, nia ya kupendeza
43	**Rapacious**	**Eyelash**	greedy, selfish	tamaa, ubinafsi
44	**Specious**	**Umbrella**	Something that sounds reasonable but is not real or true	Kitu ambacho kinaonekana kuwa kizuri lakini si kweli au kweli

Word pairs for Exp 4 continued

45	**Ostensible**	**Balcony**	something that is said to be true, but people have doubts	kitu kinachojulikana kuwa ni kweli, lakini watu wana mashaka
46	**Supercilious**	**Knuckle**	arrogant, self-important	kiburi, binafsi muhimu
47	**Unctuous**	**Grape**	full of praise and kindness, but is obviously insincere	kamili ya sifa na fadhili, lakini ni dhahiri kuwa hafifu
48	**Ephemeral**	**Berry**	lasts only for a very short time	hudumu kwa muda mfupi sana
49	**Truculent**	**Sheep**	bad-tempered, aggressive	mbaya-hasira, fujo
50	**Venerable**	**Helmet**	old and wise, deserving respect	zamani na hekima, kuheshimiwa

B.2 Coherent Text Materials for Exp 5

Experiment 5 - Coherent version of study materials

Block one

Item No	Item
1	The Moon, is a spherical, rocky body, probably with a small metallic core, that revolves around Earth in a slightly eccentric orbit at a mean distance of about 384,000 km.
2	Its equatorial radius is 1,738 km, and its shape is slightly flattened in such a way that it bulges a little in the direction of Earth.
3	Its mass distribution is not uniform, the centre of mass is displaced about 2 km toward Earth relative to the centre of the lunar sphere.
4	The Moon also has surface mass concentrations, called mascons for short, that cause the Moon's gravitational field to increase over local areas.
5	The Moon has no global magnetic field like that of Earth, but some of its surface rocks have remnant magnetism, which indicates one or more periods of magnetic activity in the past.
6	The Moon presently has very slight seismic activity and little heat flow from the interior, indications that most internal activity ceased long ago.

=158 words

Block two

1	Scientists now believe that more than four billion years ago the Moon was subject to violent heating, which resulted in its chemical separation, into a less dense crust and a denser underlying mantle.
2	The Moon's initial period of violent heating was followed hundreds of millions of years later by a second episode of heating, this time from internal radioactivity, which resulted in volcanic outpourings of lava.
3	Because of the Moon's small size and mass, its surface gravity is only about one-sixth of the Earth's; it retains so little atmosphere that the molecules of any gases present on the surface move without collision.
4	In the absence of an atmospheric shield to protect the Moon's surface from bombardment, countless bodies ranging in size from asteroids to tiny particles have struck and cratered the Moon.
5	Countless impacts on the Moon's surface have formed a debris layer, or regolith, consisting of rock fragments of all sizes down to the finest dust.
6	In the ancient past the largest impacts on the Moon's surface made great basins, some of which were later partly filled by the enormous lava floods.
7	These great dark plains, called maria (singular mare [Latin: "sea"]), are clearly visible to the naked eye from Earth.
8	The dark maria and the lighter highlands, whose unchanging patterns many people recognize as the "man in the moon," constitute the two main kinds of lunar territory.
9	The Mascons are regions on the Moon where particularly dense lavas rose up from the mantle and flooded into basins.
10	Lunar mountains, located mostly along the rims of ancient basins, are tall but not steep or sharp-peaked, because all lunar landforms have been eroded by the unending rain of impacts.

=279 words

Coherent Text Materials for Exp 5 continued

Block three

Item No	Item
1	In addition to its nearness to Earth, the Moon is relatively massive compared with the planet, with the ratio of their masses being much larger than those of other natural satellites to the planets that they orbit.
2	The Moon and Earth consequently exert a strong gravitational influence on each other, forming a system that has distinct properties and behaviour of its own.
3	Although the Moon is commonly described as orbiting Earth, it is more accurate to say that the two bodies orbit each other about a common centre of mass.
4	Called the barycentre, this point lies inside Earth about 4,700 km from its centre.
5	Also more accurately, it is the barycentre, rather than the centre of Earth, that follows an elliptical path around the Sun in accord with Kepler's laws of planetary motion.
6	The orbital geometry of the Moon, Earth, and Sun gives rise to the Moon's phases and to the phenomena of lunar and solar eclipses.
7	The Moon displays four main phases: new, first quarter, full, and last quarter.
8	New moon occurs when the Moon is between the Earth and the Sun, and thus the side of the Moon that is in shadow faces Earth.
9	Full moon occurs when the Moon is on the opposite side of Earth from the Sun, and thus the side of the Moon that is illuminated faces Earth.
10	First and last quarter of the Moon's phases, in which half the Moon appears illuminated, occur when the Moon is at a right angle with respect to the Sun when viewed from Earth.

= 257 words

Block four

1	From the perspective of a person on Earth, a solar eclipse happens when the Moon comes between the Sun and Earth, and a lunar eclipse happens when the Moon moves into the shadow of Earth cast by the Sun.
2	Solar eclipses occur at new moon, and lunar eclipses occur at full moon.
3	Eclipses do not occur every month, because the plane of the Moon's orbit is inclined to that of Earth's orbit around the Sun by about 5°, therefore at most new and full moons, the Earth, Sun, and Moon are not in a straight line.
4	The distance between the Moon and Earth varies rather widely because of the combined gravity of the Earth, the Sun, and the planets.
5	For example, in the last three decades of the 20th century, the Moon's apogee, the farthest distance that it travels from Earth in a revolution, ranged between about 404,000 and 406,700 km.
6	While its perigee, the closest that it comes to Earth, ranged between about 356,500 and 370,400 km.
7	The gravitational attraction between Earth and the Moon, have braked the Moon's spin such that it now rotates at the same rate as it revolves around Earth and thus always keeps the same side facing the planet.

= 205 words

Total word count = 899

B.3 Practice & Final Test Items Exp 5

Experiment 5 – practice test and final test items and correct answer coding

Item No	Item	Correct answer, additional accepted answers
1	The moon is thought to have a small, _____ core.	Metallic, metal
2	The moon's centre of mass is displaced about 2km toward _____.	earth
3	Surface mass concentrations on the moon cause the Moon's _____ field to be increased over local areas.	Gravitational
4	The _____ magnetism in surface rocks, indicate the moon had periods of magnetic activity in the past.	Remnant, residual, remaining, residue, trace, fragment, slight, little, low, small, weak, partial, leftover, part
5	Very slight _____ activity and little interior heat flow, indicates that the internal activity of the moon ceased long ago.	Seismic, tectonic
6	More than _____ billion years ago the Moon is thought to have been subjected to violent heating.	Four, 4
7	The initial violent heating of the moon, led to its differentiation into a less dense crust, and a denser underlying _____.	Mantle
8	The moon has approximately one _____ of the Earth's gravity.	Sixth, 6
9	A second heating of the moon from internal radiation led to outpourings of _____.	Lava, magma, molten rock
10	The moon has less _____ than the Earth due to its small mass and size.	Gravity, gravitation, gravitational
11	Due to its lack of atmosphere, the moon has formed a regolith, a regolith is a layer of _____.	Debris, rock, dust, rubble
12	In the ancient past the largest impacts on the moon's surface made great _____.	Basins, craters
13	The two main kinds of lunar territory are made up of lighter _____ and dark maria.	Highlands, mountains, hills, peaks, highground
14	Due to previous _____ the moon's mountains are not steep or sharp-peaked.	Impact, collisions, bombardments
15	The ratio of masses between the Earth and the Moon is_____ than other satellites to their planets of orbit.	Larger, greater, bigger
16	Although the Moon is commonly described as orbiting Earth, it is more accurate to say that the two bodies orbit each other about a common _____ of mass.	Centre, centrepoint

Practice & Final Test Items Exp 5 continued

17	The _____ is located around 4700km from the centre of the Earth.	Barycentre
18	In line with _____'s law of planetary motion, the barycentre forms an elliptical path around the sun.	kepler
19	There are four phases of the moon: new, first quarter, full, _____ quarter.	Last
20	The orbital geometry of the Sun, Moon and Earth give rise to the Moon's _____.	phases
21	A _____ moon occurs when the moon is between the Earth and the sun.	new
22	A _____ eclipse occurs when the moon is full.	solar
23	The moon's orbit is inclined to that of the Earth's orbit around the sun by about _____ degrees.	Five, 5
24	Due to the combined gravity of the Earth, Sun and planets, the_____ between the Moon and Earth varies widely.	Distance, space, gap
25	The moon's apogee is the _____ distance from the earth the moon achieves in one revolution.	Furthest, largest, greatest, maximum, farthest, longest, biggest
26	The moon's _____ is the shortest distance from the earth the moon achieves in one revolution.	perigee
27	The gravitational attraction between the Earth and the Moon, have _____ the Moon's spin.	Braked, reduced slowed, decreased, stopped, halted, ceased

B.4 Study Items & Test Questions Materials for Exp 6

A Summer in the Twenties

Trial No	Study Item	Test Item	Answer
1	'MORNING, MASTER TOM,' said Stevens, holding the front door. 'Don't you bother about your traps. Pennycuick and me'll get that lot in.'	Stevens says he will get in the t_____ with Pennycuick for master Tom.	traps
2	'Thanks. Where's the General?' 'In the Collection Room.'	Stevens tells Tom that the General is in the c_____ room.	collection
3	'We got your wire yesterday, so he'll be expecting you. Good journey, Master Tom?' 'Fine, thanks.'	Stevens tells master Tom he received his w_____ yesterday.	wire
4	In fact in the undiminishing daze of love Tom had barely noticed the battering French trains, or the crossing, or the somehow less heavy-breathing English engines.	The French trains are described as b_____.	battering
5	The only part of the last two days that had been free of the unreality of dream had been the evening in the Smoking Room at the United University, spent writing a nine-page letter to Judy.	Tom says the smoking room of the U_____ University, was free of the unreality of dream.	united
6	He didn't go in at once but stood under the portico looking round.	Master Tom stands under the p_____ before going into the house.	portico
7	Even Sillerby was less solid than usual.	Sillerby was described as being less s_____ than usual.	solid
8	The first faint bloom of weeds was beginning to show in the sickly rose-beds that ringed the turning-circle of gravel.	Sickly r_____ were seen to ring the turning-circle of gravel.	rose-beds
9	The paint was flaking on the billiard-room window.	In the b_____ room there was seen to be paint flaking on the window.	billiard
10	Usually these dilapidations, and the difficulty of getting them all attended to with Sillerby's diminished and increasingly arthritic staff, oppressed him.	Sillerby's staff were described as diminished and increasingly a_____.	arthritic
11	In this glittering noon the dilapidations became part of his mood, symbols of growth and of transience, of the need to snatch the instant.	The dilapidations were symbols of growth, transience and the need to snatch the i_____.	instant
12	'Any news of Master Gerald?' he said casually.	Master Tom asked if there was any news of Master G_____.	gerald

Study Items & Test Questions Materials for Exp 6 continued

13	'Not that I have heard, Master Tom. Still with Miss Nan, I believe, and doing well as can be hoped.'	Master Gerald was thought to still be with Miss N_____.	nan
14	'Oh . . . Right, I'll go and find the General, That middle-size case is all laundry so it might as well go straight out to Mrs. Bird.'	The middle-sized case of l_____ was to go straight out to Mrs. Bird.	laundry
15	'And I've torn my green plus-fours, so don't hang 'em up.' 'I have a suit of the General's to go to London. I'll send the plus-fours with them.' 'Right oh.'	The General's suit and the plus f_____ were going to be sent to London.	fours
16	Climbing the stairs Tom began to realise a mild unease, almost shock, at the news that Gerald was 'doing well'.	Whilst climbing the stairs Tom began to realise a mild unease, almost s_____, at the news that Gerald was 'doing well'.	shock

The Liminal People

Trial No	Study Item	Test Item	Answer
1	"Suleiman." I find him with his family, his wife, and his two children ages three and seven.	The ages of Suleiman's children are three and s_____.	seven
2	His tastes lean toward the moderate: not a lot of foreign products in the house aside from the expansive television.	Suleiman's tastes lean toward the m_____.	moderate
3	Minus the drug running, and Suleiman would be the perfect model for the modern Morocco.	Suleiman would be a perfect model of M_____ if not for the drug running.	Morroco
4	I take my shoes off before entering his house and wave my hand at his wife, letting her know it's OK to keep the veil down.	Taggert took off his s_____ before entering the house.	shoes
5	"Taggert, say hello to my children," Suleiman commands.	Suleiman tells Taggert to say hello to his c_____.	children
6	He thinks I'm from London so he speaks with a fake Cockney accent.	Suleiman talks to Taggert in a fake c_____ accent.	cockney
7	He wants his children to speak English, so I'm put through this cross-generational farce every time I come by. I hate children.	Taggert is put through a cross-g_____ farce because Suleiman wants his children to speak English.	generational
8	Luckily, I don't have to tolerate them for much longer than it takes Suleiman's wife to make the customary tea. We are left in the kitchen alone.	Suleiman's wife makes the customary t_____.	tea
9	"Was Omar so bad?" he says, examining the scowl on my face. "He tried to swindle. The boss	O_____ is described as having tried to swindle.	Omar

Study Items & Test Questions Materials for Exp 6 continued

	will have to talk to his people; don't be surprised if the guy comes up missing."		
10	I say in rapid-fire Arabic only to be interrupted by Suleiman's brief but fervent prayer for the idiot's soul.	A_____ is spoken in rapid fire by Taggert.	Arabic
11	The rumor goes that Suleiman used to be in training a mullah before the boss got a hold of him. "This isn't about that."	The rumour says Suleiman used to be in training for a M_____.	Mullah
12	I pull out the recorder and slide it back to him. Already erased.	Taggert slides a r_____ across to Suleiman, already erased.	recorder
13	Sully looks at it suspiciously, then brings his long-scanning, desert eyes up to meet mine.	Sully looks at the recorder suspiciously, before bringing his d_____ eyes up to meet Taggert.	desert
14	"You asked me to check it once a month when you first came to us. But we haven't used that safe house for a few months now."	That s_____ house hasn't been used for a few months now.	safe
15	"I'm not mad," I lie. "I just want to know if you played it for anyone else." Has he told Nordeen?	Taggert tells Suleiman he isn't mad but just wants to know whether the recorder was p_____ for anyone else.	played
16	"I've only been home twenty minutes. I haven't even had time to see the Old Man yet," he says slowly. "If it's OK with you, I'd like to tell him about it myself."	Suleiman says he has been home for t_____ minutes.	twenty

The King's Last Song

Trial No	Study Item	Test Item	Answer
1	William is always the first awake. He lies in the dark for a few moments listening to the roosters crow.	William listens to the r_____ crow whilst lying in the dark.	rooster
2	The cries cascade across the whole floodplain, all the way to the mountains, marking how densely populated the landscape is.	The cries cascade their way across the floodplain and all the way to the m_____.	mountains
3	William is himself in those moments. At every other time of the day he is working.	William has moments when he is himself and at every other time of day he is w_____.	working
4	William looks at the moon through the open shutters.	William looks at the m_____ through the open shutters.	moon
5	The moonlight on the mosquito net breaks apart into a silver arch.	The moonlight hits the mosquito net and breaks apart into a s_____ arch.	silver

Study Items & Test Questions Materials for Exp 6 continued

6	This is his favourite moment; he uses it to think of nothing at all, but just to look. Then he rolls to his feet.	William uses his f_____ moment to think of nothing at all.	favourite
7	The house is a clock. Its shivering tells people who has got up and who will be next.	The house is described as a c_____.	clock
8	One of his cousins turns over. In the main room, William steps over the girls asleep in a row on the floor.	William steps over the g_____ asleep in a row on the floor.	girls
9	He swings down the ladder into his waiting flip-flops and pads to the kitchen shed.	William swings down the l_____ into his waiting flip-flops.	ladder
10	Embers glow in moulded rings that are part of the concrete tabletop.	Embers glow in m_____ rings that are part of the concrete tabletop.	moulded
11	William leans over, blows on the fire, feeds it twigs, and then goes outside to the water pump.	William feeds the fire t_____.	twigs
12	Candles move silently through the trees, people going to check their palm-wine stills or to relieve themselves.	C_____ move silently through the trees.	candles
13	A motorcycle putters past; William says hi. He boils water and studies by candlelight.	William says hi to a m_____ as it putters past.	motorcycle
14	He has taught himself English and French and enough German to get by. Now he is teaching himself Japanese. He needs these languages to talk to people.	William has taught himself English, French and e_____ German to get by.	enough
15	On the same shelf as the pans is an old ring binder. It is stuffed full with different kinds of paper, old school notebooks or napkins taken from restaurants.	An old r_____is found on a shelf with the pans.	ringbinder
16	Each page is about someone: their name, address, e-mail, notes about their family, their work, what they know. William has learned in his bones that survival takes the form of other people.	William has learned in his bones that s_____ takes the form of other people.	survival

The Little Winter

Trial No	Study Item	Test Item	Answer
1	At the airport, Gloria rented a car.	Gloria rented a car at the a_____.	airport

Study Items & Test Questions Materials for Exp 6 continued

2	She decided to drive until just outside Jean's town and check into a motel.	Gloria had decided to drive to just outside Jean's town and check into a m_____.	motel
3	Jean was a talker. A day with Jean would be enough. A day and a night would be too much.	Jean was described as a t_____.	talker
4	Just outside Jean's town was a monastery where the monks raised dogs. Maybe she would find her dog there tomorrow.	Gloria was to go to a monastery where monks raised d_____.	dogs
5	She would go over to the monastery early in the morning and spend the rest of the day with Jean. But that was it, other than that, there wasn't much of a plan.	Gloria was p_____ to spend the rest of the day with Jean after visiting the monastery.	planning
6	The day was cloudy and there was a great deal of traffic. The land falling back from the highway was green and still.	The land falling back from the h_____ was described as green and still.	highway
7	It seemed to her a slightly morbid landscape, obelisks and cemeteries, thick drooping forests, the evergreens dying from the top down.	The landscape seemed to Gloria to be slightly m_____.	morbid
8	Of course there was hardly any place to live these days. A winding old road ran parallel to the highway and Gloria turned off and drove along it until she came to a group of cabins.	Gloria turned off the road a drove until she came to a group of c_____.	cabins
9	The cabins were white with little porches but the office was in a structure built to resemble a tepee.	The office was in a s_____ built to resemble a tepee.	structure
10	There was a dilapidated miniature golf course and a wooden tower from the top of which you could see into three states.	You could see into three states from the top of the w_____ tower.	wooden
11	But the tower leaned and the handrail curving optimistically upward was splintered and warped, and only five steps from the ground a rusted chain prevented further ascension. Gloria liked places like this.	Only f_____ steps from the ground there was a rusted chain that prevented further ascension.	five
12	In the tepee, a woman in a housedress stood behind a pink formica counter.	The woman in the housedress stood behind a pink f_____ counter.	formica
13	A glass hummingbird coated with greasy dust hung in one window.	Hanging in one of the windows was a glass h_____ coated in dust.	hummingbird

Study Items & Test Questions Materials for Exp 6 continued

14	Gloria could smell meatloaf cooking.	Gloria could smell m_____ cooking.	meatloaf
15	The woman had red cheeks and white hair, and she greeted Gloria extravagantly, but as soon as Gloria paid for her cabin she became morose.	The woman greeted Gloria e_____, before she became morose.	extravagently
16	She gazed at Gloria glumly as though perceiving her as one who had already walked off with the blankets, the lamp and the painting of the waterfall.	Gloria thought the woman had already perceived her as walking off with the painting of the w_____.	waterfall

Appendix C

Chapter Four Materials

C.1 Study Materials for Exp 7

Items for Penguins				
Trial Number	**Full study item**	**Simple Study Item**	**Simple Test**	**Follow Up Elaborate Q**
1	Today, wild penguins exhibit no particular fear of human tourists, this is because they are not used to danger from animals on solid ground.	Today, wild penguins exhibit no particular fear of human tourists.	Today, wild penguins exhibit no particular fear of human WHAT?	Why is that?
2	Penguins can filter out the ocean water from their bloodstream, they need to do this because they ingest a lot of seawater while hunting for fish.	Penguins can filter out the ocean water from their bloodstream.	Penguins can filter out the ocean water from their WHAT?	Why is that useful?
3	While swimming, penguins maintain a defence against predation by the black and white colouring on their body. Their colouring blends in with the sea from above and it blends in with the sky from below.	While swimming, penguins maintain a defence against predation by the black and white colouring on their body.	While swimming, penguins maintain a defence against predation by the black and white WHAT on their body?	Why is that?
4	Penguins trap a layer of air close to their skin with their dense plumage, this helps them with buoyancy and heat conservation in the water.	Penguins trap a layer of air close to their skin with their dense plumage.	Penguins trap a layer of air close to their skin with their dense WHAT?	Why is that useful?
5	During a dive penguins have the ability to greatly slow their resting heartrate, this is	During a dive penguins have the ability to greatly slow	During a dive penguins have the ability to greatly	Why is that useful?

Study Materials for Exp 7 continued

	so that they can conserve energy whilst hunting for food.	their resting heartrate.	slow their resting WHAT?	
6	After a season of hunting, penguins spend two to three weeks on land, this is because they undergo what is called the catastrophic molt.	After a season of hunting, penguins spend two to three weeks on land.	After a season of hunting, penguins spend two to three weeks on WHAT?	Why is that?
7	Penguins molt to replace the feathers of their much needed waterproof coat, because it becomes less effective after a season of hunting.	Penguins molt so that they can replace the feathers of their much needed waterproof coat.	Penguins molt so that they can replace the feathers of their much needed waterproof WHAT?	Why is that useful?
8	Amazingly, male penguins can fast for around 100 days. This is so that they can survive when no food is available.	Amazingly, male penguins can fast for around 100 days.	Amazingly, male penguins can fast for around 100 WHAT?	Why is that useful?
9	Penguins breed during the antarctic winter, so that their offspring reach independence in summer when more food is available.	Penguins breed during the antarctic winter.	Penguins breed during the antarctic WHAT?	Why is that useful?
10	Penguins have the skill to lean back and balance on their short stiff tails. This helps reduce the amount of heat lost through their feet to the ground.	Penguins have the skill to lean back and balance on their short stiff tails.	Penguins have the skill to lean back and balance on their short stiff WHAT?	Why it that useful?

Study Materials for Exp 7 continued

11	Male adult penguins will form a huddle during the coldest months, this is another way they avoid heat loss.	Male adult penguins will form a huddle during the coldest months.	Male adult penguins will huddle together during the coldest WHAT?	Why is that useful?
12	Penguins are able to control the amount of blood flow to their extremities, this reduces the amount of blood that gets cold and keeps them from freezing.	Penguins are able to control the amount of blood flow to their extremities.	Penguins are able to control the amount of blood flow to their WHAT?	Why is that useful?
		Items for Sharks		
Trial Number	**Full study item**	**Simple Study Item**	**Simple Test**	**Follow Up Elaborate Q**
13	Some sharks must swim constantly to force oxygen rich water over their gills, because they cannot naturally pump water over their gills.	Some sharks must swim constantly to force oxygen rich water over their gills.	Some sharks must swim constantly to force oxygen rich water over their WHAT?	Why is that?
14	Even whilst sleeping, sharks that must swim constantly can sustain their swimming motion, this is because the spinal cord rather than the brain controls swimming motion.	Even whilst asleep, sharks that must swim constantly can sustain their swimming motion.	Even whilst asleep, sharks that must swim constantly can sustain their swimming WHAT?	Why is that?

Study Materials for Exp 7 continued

15	Sharks only feel the need to kill when they have emptied their oil stores, because they can live off the energy from their oil stores for a long time.	Sharks only feel the need to kill when they have emptied their oil stores.	Sharks only feel the need to kill when they have emptied their oil WHAT?	Why is that?
16	Sharks benefit from having large livers full of low-density oils, this is especially useful for making them bouyant in the water.	Sharks benefit from having large livers full of low-density oils.	Sharks benefit from having large livers full of low-density WHAT?	Why is that useful?
17	Sharks have lightweight skeletons made of cartilage, as it is lighter than bone it helps to save them energy in the water.	Sharks have lightweight skeletons made of cartilage.	Sharks have lightweight skeletons made of WHAT?	Why is that useful?
18	Sharks can dislocate their jaws because they are not attached to their craniums, this is helpful when sharks attempt to kill something large.	Sharks can dislocate their jaws because they are not attached to their craniums.	Sharks can dislocate their jaws because they are not attached to their WHAT?	Why is that useful?
19	Sharks can thrust their stomachs out of their mouths, they do this to get rid of something they can't digest.	Sharks can thrust their stomachs out of their mouths.	Sharks can thrust their stomachs out of their WHAT?	Why is that useful?

Study Materials for Exp 7 continued

20	Positioned on the sides of their bodies sharks have lateral line organs, these help them detect small movements in the water.	Positioned on the sides of their bodies sharks have lateral line organs.	Positioned on the sides of their bodies sharks have lateral line WHAT?	Why is that useful?
21	To locate their prey in the water sharks are also able to use electrical signals, because of this sharks sometimes mistakenly attack metal objects thinking it is prey.	To locate their prey in the water sharks are also able to use electrical signals.	To locate their prey in the water sharks are also able to use WHAT signals?	Why isn't that useful?
22	Sharks' skin is enveloped in tiny teeth, this makes it both extremely tough and hydrodynamic as the teeth direct water efficiently across the skin surface.	Sharks' skin is enveloped in tiny teeth.	Sharks' skin is enveloped in tiny WHAT?	Why is that useful?
23	Female shark skin has evolved to be three times thicker than their male counterpart, which protects them when being bitten by male sharks during mating rituals.	Female shark skin has evolved to be three times thicker than their male counterpart.	Female shark skin has evolved to be three times thicker than their male WHAT?	Why is that useful?
24	Female sharks lose their appetites around the time of birth, this is due to a biological mechanism that	Female sharks lose their appetites around the time of birth.	Female sharks lose their appetites around the time of WHAT?	Why is that useful?

Study Materials for Exp 7 continued

	helps them avoid eating their own pups.			
Items for Crocodiles				
Trial Number	**Full study item**	**Simple Study Item**	**Simple Test**	**Follow Up Elaborate Q**
25	Crocodile skin consists of a bony structure that makes it bulletproof, this helps protect it when fighting with other animals.	Crocodile skin consists of a bony structure that makes it bulletproof.	Crocodile skin consists of a bony structure that makes it WHAT?	Why is that useful?
26	Crocodiles possess excellent senses all-round, including auditory, visual and olfactory, this makes them extremely good night predators.	Crocodiles possess excellent senses all-round, including auditory, visual and olfactory.	Crocodiles possess excellent senses all-round, including auditory, visual and WHAT?	Why is that useful?
27	Saltwater crocodiles sleep with one eye open, so they can be on alert for any danger nearby.	Saltwater crocodiles sleep with one eye open.	Saltwater crocodiles sleep with WHAT eye open?	Why is that useful?
28	By adulthood crocodiles have developed long and streamlined bodies, to help them move quickly through the water.	By adulthood crocodiles have developed long and streamlined bodies.	By adulthood crocodiles have developed long and streamlined WHAT?	Why is that useful?

Study Materials for Exp 7 continued

29	Crocodiles have incredibly powerful muscles for closing their jaws, so that once they catch their prey it can't escape.	Crocodiles have incredibly powerful muscles for closing their jaws.	Crocodiles have incredibly powerful muscles for closing their WHAT?	Why is that useful?
30	Once crocodiles have captured their prey, they perform a death roll. Forcefully dragging their prey under water helps to separate the limbs from the bodies.	Once crocodiles have captured their prey, they perform a death roll.	Once crocodiles have captured their prey, they perform a death WHAT?	Why is that useful?
31	Instead of chewing their food crocodiles swallow stones, this helps them to break down the food inside their stomachs.	Instead of chewing their food crocodiles swallow stones.	Instead of chewing their food crocodiles swallow WHAT?	Why is that useful?
32	Crocodiles are able to digest all elements of their prey including, bones, hooves, horns and shells. This is because they have the most acidic stomach of any vertebrate.	Crocodiles are able to digest all elements of their prey including, bones, hooves, horns and shells.	Crocodiles are able to digest all elements of their prey including, bones, hooves, horns and WHAT?	Why is that?
33	When crocodiles eat their catch they appear to produce tears, air coming into contact with their tear glands while	When crocodiles eat their catch they appear to produce tears.	When crocodiles eat their catch they appear to produce WHAT?	Why is that?

Study Materials for Exp 7 continued

	they eat, forces tears to flow.			
34	Crocodiles can survive in a state of not eating for over a year, having an efficient metabolism means they can store nearly all the food they consume.	Crocodiles can survive in a state of not eating for over a year.	Crocodiles can survive in a state of not eating for over a WHAT?	Why is that?
35	Crocodiles are often seen on shore with their mouths open to release heat, they need to do this because they do not have sweat glands.	Crocodiles are often seen on shore with their mouths open to release heat.	Crocodiles are often seen on shore with their mouths open to release WHAT?	Why is that?
36	Crocodiles tongues have limited movement due to being held in place by a membrane, this helps them to avoid biting it when they clamp their jaws closed around their prey.	Crocodiles tongues have limited movement due to being held in place by a membrane.	Crocodiles tongues have limited movement due to being held in place by a WHAT?	Why is that?

C.2 Practice Test Materials for Exps 8 & 9

Items for Penguins		
Trial Number	**Exp 8 test practice**	**Exp 9 test practice**
1	Today, wild penguins exhibit no WHAT fear of human tourists?	Today, wild penguins exhibit no particular WHAT of human tourists?
2	Penguins can filter out the WHAT water from their bloodstream?	Penguins can WHAT out the ocean water from their bloodstream?
3	While swimming, penguins maintain a WHAT against predation by the black and white colouring on their body?	While swimming, penguins WHAT a defence against predation by the black and white colouring on their body?
4	Penguins trap a WHAT of air close to their skin with their dense plumage?	Penguins WHAT to a layer of air close to their skin in their dense plumage?
5	During a dive penguins have the WHAT to greatly slow their resting heartrate?	During a dive penguins have the ability to greatly WHAT their resting heartrate?
6	After a WHAT of hunting, penguins spend two to three weeks on land?	After a season of hunting, Penguins WHAT two to three weeks on land?
7	Penguins molt so that they can replace the WHAT of their much needed waterproof coat?	Penguins molt so that they can WHAT the feathers of their much needed waterproof feather coat?
8	Amazingly, WHAT penguins can fast for around 100 days?	Amazingly, male penguins can WHAT for around 100 days?
9	Penguins breed during the WHAT winter?	Penguins WHAT during the winter?
10	Penguins have the WHAT to lean back and balance on their short stiff tails?	Penguins are able to lean back and WHAT on their short stiff tails?
11	Male adult penguins will form a WHAT during the coldest months?	Male adult penguins will WHAT a huddle during the coldest months?
12	Penguins are able to control the WHAT of blood flow to their extremities?	Penguins are able to WHAT to the amount of blood flow to their extremities?

Practice Test Materials for Exps 8 & 9 continued

Items for Sharks		
Trial Number	**Exp 8 test practice**	**Exp 9 test practice**
13	Some sharks must swim constantly to force WHAT rich water over their gills?	Some sharks must swim constantly to force oxygen rich water over their gills.
14	Even whilst WHAT, sharks that must swim constantly can sustain their swimming motion?	Even whilst asleep, sharks that must swim constantly can sustain their swimming motion.
15	Sharks only feel the WHAT to kill when they have emptied their oil stores?	Sharks only feel the need to kill when they have emptied their oil stores.
16	Sharks benefit from having large WHAT full of low-density oils?	Sharks benefit from having large livers full of low-density oils.
17	Sharks have lightweight WHAT made of cartilage?	Sharks have lightweight skeletons made of cartilage.
18	Sharks can dislocate their WHAT because they are not attached to their craniums?	Sharks can dislocate their jaws because they are not attached to their craniums.
19	Sharks can thrust their WHAT out of their mouths?	Sharks can thrust their stomachs out of their mouths.
20	Positioned on the sides of their WHAT sharks have lateral line organs?	Positioned on the sides of their bodies sharks have lateral line organs.
21	To locate their WHAT in the water sharks are also able to use electrical signals?	To locate their prey in the water sharks are also able to use electrical signals.
22	Sharks' WHAT is enveloped in tiny teeth?	Sharks' skin is enveloped in tiny teeth.
23	Female shark skin has evolved to be WHAT times thicker than their male counterpart?	Female shark skin has evolved to be three times thicker than their male counterpart.
24	Female sharks lose their WHAT around the time of birth?	Female sharks lose their appetites around the time of birth.

Practice Test Materials for Exps 8 & 9 continued

Items for Crocodiles		
Trial Number	**Exp 8 test practice**	**Exp 9 test practice**
25	Crocodile skin consists of a bony WHAT that makes it bulletproof?	Some sharks must WHAT constantly to force oxygen rich water over their gills?
26	Crocodiles possess excellent WHAT all-round, including auditory, visual and olfactory?	Even whilst asleep, sharks that must swim WHAT can sustain their swimming motion?
27	WHAT crocodiles sleep with one eye open?	Sharks only feel the need to WHAT when they have emptied their oil what?
28	By WHAT Crocodiles have developed long and streamlined bodies?	Sharks WHAT from having large livers full of low-density oils?
29	Crocodiles have incredibly powerful WHAT for closing their jaws?	Sharks have lightweight skeletons WHAT of cartilage?
30	Once crocodiles have captured their WHAT, they perform a death roll?	Sharks can WHAT their jaws because they are not attached to their craniums?
31	Instead of chewing their WHAT crocodiles swallow stones?	Sharks can WHAT their stomachs out of their mouths?
32	Crocodiles are able to digest all WHAT of their prey including, bones, hooves, horns and shells?	WHAT on the sides of their bodies sharks have lateral line organs?
33	When crocodiles eat their WHAT appear to produce tears?	To WHAT their prey in the water, sharks are also able to use electrical signals?
34	Crocodiles can survive in a WHAT of not eating for over a year?	Sharks' skin is WHAT in tiny teeth?
35	Crocodiles are often seen on WHAT with their mouths open to release heat?	Female shark skin has WHAT to be three times thicker than their male counterpart?
36	Crocodiles tongues have limited WHAT due to being held in place by a membrane?	Female sharks WHAT their appetites around the time of birth?

C.3 Study & Test Materials for Exp 10

Pair No	Coffee Fact Pair	Practice Test Item: simple (1), elab (1 + 2)	First Item Correct	Second Item Correct
1	Coffee was first discovered when a goatherder witnessed increased energy in his herd. He realised his herd had been eating the fruit of the coffee plant.	Coffee was first discovered when a g_____ witnessed increased energy in his herd (1). He realised his herd had been eating the fruit of the coffee p_____ (2).	goatherder	plant
2	Coffee was later consumed as a food by mixing the coffee beans with animal fat. This mixture created a high energy snack that was eaten by early African tribes.	Coffee was first consumed as a food by mixing the beans with animal f_____ (1). This mixture created a high energy snack that was eaten by early African t_____ (2).	fat	tribes
3	As early as the thirteenth century Muslims were drinking coffee. Presumably even before this time it had been developed into a hot drink.	As early as the t_____ century muslims were drinking coffee (1). Presumably even before this time it had been developed into a h_____ drink (2).	thirteenth	hot
4	Coffee was only found to naturally grow in Arabia and Africa. Coffee grew in only two regions until the 1600s.	Coffee was only found to naturally grow in Arabia and A_____ (1). Coffee grew in only two regions until the 1_____(2).	africa	1600s

Study & Test Materials for Exp 10 continued

5	Coffee was thought to be introduced to India by a smuggler named Baba Budan. Baba Budan left Mecca with fertile seeds strapped to his chest.	Coffee was thought to be introduced to I_____ by a smuggler named Baba Budan (1). Baba Budan left Mecca with fertile seeds strapped to his c_____ (2).	india	chest
6	The Dutch founded a coffee estate on the island of Java in 1616. Java has become synonymous with coffee and is still grown there today.	The D_____ founded a coffee estate on the island of Java in 1616 (2). Java has become synonymous with coffee and is still grown there t_____ (1).	dutch	today
7	Coffee first crossed the Atlantic in 1727. Coffee was thought to be smuggled into Brasil by a spy.	Coffee first crossed the A_____ in 1727 (1). Coffee was thought to smuggled into B_____ by a spy (2).	atlantic	brasil
8	Hawaii is the only state in the USA to grow coffee. Its Kona coffee is grown on volcanic mountains.	H_____ is the only state in the USA to grow coffee (1). Its Kona coffee is grown on v_____ mountains (2).	hawaii	volcanic
9	A coffee tree takes around three to four years after being planted to become productive. A productive tree also occurs around a year after white blossoms show.	A coffee tree takes around t_____ to four years after being planted to become productive (1). A productive tree also occurs around a year after w_____ blossoms show (2).	three	white

Study & Test Materials for Exp 10 continued

10	Once mature the trees will continually produce coffee cherries. Continual production means both ripe and unripe cherries are always present.	Once m_____ the trees will continually produce coffee cherries (1). Continual production means both ripe and u_____ cherries are always present (2).	mature	unripe
11	The two main varieties of coffee are Arabica and Robusta. The soil, altitude and climate can all impact the coffee's flavour.	The two main varieties of coffee are Arabica and R_____ (1). The soil, altitute and c_____ can all impact the coffee's flavour (2).	robusta	climate
12	Arabica coffee is descended from Ethiopian coffee trees. Arabica coffee is mild and aromatic.	Arabica coffee is descended from E_____ coffee trees (1). Arabica coffee is mild and a_____ (2).	ethiopia	aromatic
13	Arabica is grown best at high altitudes and mild temperatures. Arabica accounts for 70% of the world coffee market.	Arabica is grown best at high altitudes and mild t_____ (1). Arabica accounts for 70% of the world coffee m_____(2).	temperatur es	market
14	Robusta coffee trees can be found in areas of Southeast Asia and Brasil. Robusta trees produce a bitter and more caffeinated coffee than Arabica.	Robusta coffee trees can be found in areas of Southeast A_____ and Brasil (1). Robusta trees produce a bitter and more c_____ coffee than Arabica (2).	asia	caffeinated
15	The Robusta plant can thrive at higher temperatures and lower altitudes than Arabica. Surviving in these conditions results in it being more robust	The Robusta plant can thrive at higher temperatures and l_____ altitudes than arabica (1). Surviving in these conditions results in it being more robust as the	lower	name

Study & Test Materials for Exp 10 continued

	as the name suggests.	n_____ suggests (2).		
16	The Robusta bean is smaller and rounder than the arabica bean. Although robust its bitter flavour means it only represents 30% of the world coffee market.	The Robusta bean is smaller and r_____ than the arabica bean (1). Although robust its b_____ flavour means it only represents 30% of the world coffee market (2).	rounder	bitter
17	A coffee bean is actually more of a cherry like fruit. The cherries turn red when ripe for picking.	A coffee bean is actually more of a cherry like f_____ (1). The cherries turn r_____ when ripe for picking (2).	fruit	red
18	The skin of the coffee cherry is thick and bitter. This thick and bitter layer is called the exocarp.	The s_____ of the coffee cherry is thick and bitter (1). The thick and bitter layer is called the e_____(2).	skin	exocarp.
19	The fruit beneath the skin is sweet and grape like in texture. The sweet layer is called the mesocarp.	The fruit beneath the skin is sweet and g_____ like in texture (1). The sweet layer is called the m_____ (2).	grape	the mesocarp.
20	The beans are protected by a slimy honey layer. The slimy layer is termed the parenchyma.	The beans are protected by a slimy h_____ layer (1). The slimy layer is termed the p_____(2).	honey	parenchyma.

Study & Test Materials for Exp 10 continued

21	Finally the beans are covered by a parchment like envelope. The envelope covering is called the endocarp.	Finally the beans are covered by a p_____ like envelope (1). The envelope covering is called the e_____ (2).	parchment	endocarp
22	Inside the envelope there are two bluish-green coffee beans. Covering the bluish-green coffee beans is a membrane called the spermoderm.	Inside the envelope there are t_____ bluish-green coffee beans (1). Covering the bluish-green coffee beans is a membrane called the s_____ (2).	two	spermoderm
23	Northern regions harvest once a year between September and March. Whereas Southern regions will harvest once a year between April and May.	Northern regions harvest once a year between September and M_____ (1). Whereas Southern regions will harvest once a year between A_____ and May (2).	march	april
24	Harvesting involves stripping the whole branch by hand. Selectively picking in this way is more expensive and is reserved for Arabica beans.	Harvesting involves stripping the whole b_____ by hand (1). Selectively picking in this way is more expensive and is reserved for A_____ beans (2).	branch	arabica
25	Beans must be processed straight away either by wet method or dry method. For processing, the object of this initial step is to dry out the beans.	Beans must be processed straight away either by w_____ method or dry method (1). For processing, the o_____ of this initial step is to dry out the beans (2).	wet	object

Study & Test Materials for Exp 10 continued

26	The dry method involves drying the cherries in sunlight. The cherries are periodically raked to rotate them for around 7-10 days.	The dry method involves drying the cherries in s_____ (1). The cherries are periodically r_____ to rotate them for around 7-10 days (2).	sunlight	raked
27	The wet method involves washing the pulp from the cherries via machine. Following washing, the beans are then dried by sunlight or dryers.	The wet method involves washing the p_____ from the cherries via machine (1). Following washing, the beans are then dried by sunlight or d_____ (2).	pulp	dryers
28	Once dried beans are hulled to remove any remaining layers. Then the beans can be graded based on size and density.	Once dried beans are h_____ to remove any remaing layers (1). Then the beans can be graded based on size and d_____ (2).	hulled	density
29	Coffee is then shipped unroasted in bags of jute or sisal. Due to the bean's colour, the shipped coffee is called green coffee.	Coffee is then shipped u_____ in bags of jute or sisal (1). Due to the bean's colour, the shipped coffee is called g_____ coffee (2).	unroasted	green
30	The green beans are then roasted in large drums at about 288 degrees celcius. The turning motion of the roaster keeps the beans from burning.	The green beans are roasted in large d_____ at about 288 degrees celcius (1). The turning motion of the roaster keeps the beans from b_____ (2).	drums	burning

Study & Test Materials for Exp 10 continued

31	After eight minutes of roasting the beans will make a popping sound and double in size. After eight minutes the beans have reached 204 degrees celsius.	After eight minutes of roasting the beans will make a p_____ sound and double in size (1). After eight minutes the beans have r_____ 204 degrees celsius (2).	popping	reached
32	After they have doubled in size, the beans then start to brown and release oils. The substance released as the beans start to brown is known as coffee essence or caffeol.	After they have doubled in size, the beans then start to brown and release o_____ (1). The substance released as the beans start to brown is known as coffee e_____ or caffeol (2).	oils	essence
33	A second pop occurs between three and five minutes later. The second pop signals that the bean is fully roasted.	A second pop occurs between three and f_____ minutes later (1). The second pop signals that the bean is fully r_____ (2).	five	roasted
34	After seven minutes you will have a roast for distribution to the mass-market. The roast at seven minutes is called lightly roasted.	After seven minutes you will have a roast for d_____ to the mass-market (1). The roast at seven minutes is called l_____ roasted (2).	seven	distribution
35	After around ten minutes you will have a more full bodied roast. The roast at ten minutes is known as a medium roast.	After around ten minutes you will have a more full b_____ roast (1). The roast at ten minutes is known as m_____ roast (2).	bodied	medium

Study & Test Materials for Exp 10 continued

36	At around twelve minutes you have a French or Viennese coffee. The roast has developed after twelve minutes into a dark roast.	At around twelve minutes you have a F_____ or Viennese coffee (1). The roast has d_____ after twelve minutes into a dark roast (2).	French	developed
37	Finally when the beans start to smoke at fourteen minutes you get espresso roast. Espresso roast is the darkest roast available.	Finally when the beans start to s_____ at fourteen minutes you get espresso roast (1). Espresso roast is the d_____ roast available (2).	smoke	darkest
38	There is a Madagascan coffee species that naturally produces decaffeinated beans. The Madagascan bean is the only coffee bean of its kind.	There is a Madagascan coffee species that naturally produces d_____ beans (1). The Madagascan bean is the o_____ coffee bean of its kind (2).	decaffeinated	only
39	Yet, decaffeinated coffee can be made in two ways from regular beans. Decaffeinated coffee can be achieved by washing the caffeine out of the bean before roasting.	Yet, decaffeinated coffee can be made in t_____ ways from regular beans (1). Decaffeinated coffee can be achieved by w_____ the caffeine out of the bean before roasting (2).	two	washing
40	One decaffeinating method uses a chemical solvent. This solvent substance is washed over the beans before they are dried.	One decaffeinating method uses a c_____ solvent (1). This solvent substance is washed over the beans before they are d_____ (2).	chemical	dried

Study & Test Materials for Exp 10 continued

41	The second method uses a steam wash which enables the outer layers to be scraped away. These layers contain the most caffeine.	The second decaffeinating method uses a s_____ wash which enables the outer layers to be scraped away (1). The outer layers of the bean contain the most c_____ (2).	steam	caffeine
42	Coffee is served up in many unique ways around the world. World differences in coffee tastes are based on the type of roast, brewing methods and combined ingredients.	Coffee is served up in many u_____ ways around the world (1). World differences in coffee tastes are based on the type of roast, b_____ methods and combined ingredients (2).	unique	brewing
43	In America the preferred drink is a light roast with added cream and sugar. Although popularity of darker roasts have increased in the US with the introduction of coffee chains.	In America the preferred drink is a l_____ roast with added cream and sugar (1). Although popularity of darker roasts have increased in the US with the introduction of coffee c_____ (2).	light	chains
44	In Austria they blend two-thirds dark roast beans and one-third regular roast. The Austrian blend is known as the Viennese roast.	In Austria they blend two-thirds d_____ roast beans and one-third regular roast (1). The Austrian blend is known as the V_____ roast (2).	dark	blend
45	Espresso is brewed by forcing steam through finely ground dark roast beans. Espresso is the coffee of choice for Italy.	Espresso is brewed by f_____ steam through finely ground dark roast beans (1). Espresso is the coffee of choice for l_____ (2).	forcing	italy

Study & Test Materials for Exp 10 continued

46	Turkish coffee is more finely ground than espresso and is brewed in pots. Turkish coffee is also commonly spiced with cardamom, chicory and coriander.	Turkish coffee is more finely ground than espresso and is brewed in p_____ (1). Turkish coffee is also commonly spiced with cardamom, c_____ and coriander (2).	pots	chicory
47	Cuban coffee is extremely strong and is drunk as a shot and not sipped. Cuban coffee is typically served at the end of a meal.	Cuban coffee is extremely strong and is drunk as a is shot and not s_____ (1). Cuban coffee is typically served at the end of a m_____ (2).	sipped	meal
48	Thai coffee is also strong and spiced with chicory and sweetened with condensed milk. Thai coffee is also commonly served with ice.	Thai coffee is also strong and spiced with chicory and sweetened with c_____ milk (1). Thai coffee is also commonly served with i_____ (2).	condensed	ice