2022

Contributions of Source-Constrained Search and Late Monitoring to Recall Accuracy

Randle, James

UNIVERSITY OF
PLYMOUTH

Contributions of Source-Constrained Search and Late Monitoring to Recall Accuracy

By

James Randle

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

Doctor of Philosophy

School of Psychology

March 2022

Acknowledgements

Firstly I would like to express my immense gratitude for my supervisors, Professor Tim Hollins and Dr Michael Verde for their invaluable insight, guidance and patience throughout the entire PhD process. I have grown so much in confidence and ability as a researcher under their guidance and I could not have asked for better supervisors. Finally I want to thank my parents for their love and moral support throughout life, and for giving me the confidence to achieve my goals. Without you none of this would have been possible. Love you always.

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 without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

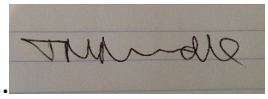
This study was financed with the aid of a studentship from the University of Plymouth School of Psychology.

A programme of advanced study was undertaken, which included a taught module, Communication of Research for Psychology, PSY555.

Conference presentations:

Randle, J.M., Hollins, T.J. & Verde, M. (2020). Joint contributions of constrained search and late monitoring to recall accuracy. Paper presented at the Experimental Psychology Society Meeting, London, January

Word count for main body of the thesis: 77636

Signed…… 

Date: 22/03/2022

Abstract

James Randle

Contributions of Source-Constrained Search and Late Monitoring to recall accuracy.

This thesis aimed to investigate in detail the processes underlying the control of recall accuracy. It is believed that recall accuracy control comprises processes occurring at memory search and post-retrieval. There is a large body of research concerning post-retrieval monitoring processes; however, less is known about how memory search is constrained and what factors affect this process.

Chapter 2 developed and tested a new paradigm for measuring the accuracy of constrained search and monitoring processes simultaneously. Participants were able to selectively retrieve one of two lists irrespective of which list was the target list, indicating successful constrained search. A key role for context reinstatement in constraining search was established through source clustering. Participants were also highly accurate in monitoring the correctness of retrieved items.

Chapter 3 required participants to constrain search to one of two sources within a single list of items (Mixed-lists). Participants were able to do this, although search accuracy was poorer than for List membership (Chapter 2). Despite this, monitoring accuracy did not differ between List membership and Mixed-lists. Only source monitoring was sensitive to source manipulations within a single experiment.

Chapter 4 presented an alternative method of measuring constrained search, which relies on curve fitting of recall latencies to estimate the size of a participant's search set. This method successfully reproduced the findings from Experiments 2.3

(List membership) and 3.2 (Mixed-lists). Unfortunately due to poor curve fits, attempts to replicate findings from Experiment 3.1 were unsuccessful.

Chapter 5 presents a retrieval model which attempts to replicate the typical pattern of declining search accuracy as the recall period progresses, with the assumption that participants' ability to search for targets does not decline throughout the recall period. This model was able to produce accuracy curves which fit the search accuracy data fairly well; however, predictions for other recall metrics were poor.

On the whole, this research demonstrates that in order to constrain search, participants must reinstate the context of the target source at retrieval, and that the success of constrained search is dependent upon the type of context or source which is encoded. Source monitoring data were generally in line with the predictions of the Source Monitoring Framework (Johnson et al., 1993).

Contents

List of tables

List of figures

Chapter 1: Thesis overview, literature review and methods

1.1 - Thesis Overview

The main aim of the current thesis is to investigate in detail how we control the accuracy of our memories. In an experimental setting researchers often require participants to learn and recall lists of items. In order to recall the items, the participant must retrieve a single instance of an item (e.g. the word “table”) from among all other encounters they may have had with that item in their lifetime. First the correct retrieval cue must be located in order to constrain memory search so that only items from the list in question come to mind, and not items from other lists, instructions or other recent encounters. Then every item that comes to mind is subjected to a monitoring process whereby it is checked for correctness. Only items that pass the correctness criterion set by the participant will be overtly output, while those which do not are withheld. This monitoring process has been extensively researched; however, we still know very little about how and to what extent we can use source (the external and intrinsic properties of an item), to constrain our search to correct information. I will particularly focus on how source constrained search and source monitoring jointly contribute to the control of memory accuracy as the retrieval process unfolds over time.

I will begin by discussing the current state of the literature, detailing how this thesis expands upon present knowledge. My first empirical chapter details a series of three experiments, centred around establishing and testing the viability of a new paradigm for measuring the joint contributions of constrained search and source

monitoring to recall accuracy. This will involve investigating the extent to which we are able to constrain search to one of two lists, and subsequently monitor the output. In addition, these experiments attempt to relate this ability to constrain search to clustering. This is a robust feature of free recall whereby the participant's output order reflects common source, temporal and semantic features among list items. This will indicate whether the new paradigm actually indexes the kind of search processes we wish to measure in source-constrained search.

The second empirical chapter explores source manipulations occurring within a single list using a novel procedure based on Externalised-Free Recall (EFR). This will give a more authentic insight into constraining search by source, as constrained search accuracy is not conflated with the ability to use inter-item associations to chain through a list. I expand on this by testing predictions from retrieval models and accuracy frameworks.

The third empirical chapter addresses the first process in source constrained retrieval; reduction of the size of the search set. Prior to retrieval, participants attempt to reduce the size of their memory search in order that they search as few incorrect items as possible. Using a mathematical approach based on modelling recall latencies I gain estimates of search set size, and contrast these between various sources to better understand set size reduction prior to search, with the intention of comparing and contrasting the two alternate methods (latency based vs EFR).

The final empirical chapter aims to expand on the previous two chapters by developing a simple model of the constrained search data. It stands to reason that constraining search will become more challenging as there are fewer correct items (targets) to retrieve. The models built attempt to correct for baseline levels of targets

changing during the recall period, providing an estimate of search efficiency in addition to accuracy.

1.2 - Review of the literature

1.2.1 - *Search processes in recall and context*

Research into search processes in free recall and the role of context has a long history. Early studies focused generally on the phenomenon of semantic clustering (Bousfield, 1953; Cofer et al., 1966). This can be described as a striking disparity between the serial order of items at study and the participant's recall output. Generally in these studies participants study lists of randomly ordered words drawn from a number of semantic categories, for instance animals and household items. On examining the participant's recall output it can be seen that recalled items appear to be organised by semantic category (context). Clustering is a useful place to start in reviewing the literature, as it will indicate the kinds of contexts which serve as successful retrieval cues during search. If memory can be organised by a particular context, then there is a good chance that this context can be used as a cue to search memory.

A landmark study was conducted by Bousfield (1953). In this experiment, participants were presented with a list of sixty randomly arranged nouns comprising fifteen exemplars of four semantic categories. They were then instructed to serially recall as many items as they could remember. To quantify clustering, Bousfield devised an empirical measure which was termed the Ratio of Repetition (RR). This is a simple ratio of the number of observed same category transitions in the subject's recall output, to the maximum possible number of same category transitions given the total number of items recalled. In this case same category transition refers to consecutive recall of two items from the same category, in essence a cluster.

In order to observe whether clustering exhibited by the subjects as measured by the RR was greater than chance, an artificial experiment was conducted. One hundred simulated recall output sequences were generated randomly without replacement and matched for recall output length with the real subjects. RR scores were calculated for each of these simulated sequences, and group means calculated for the artificial experiment and the real subjects. RR scores were 0.24 for the artificial experiment and 0.45 for the real subjects, indicating that the latter group exhibited almost double the magnitude of clustering that one would expect if clustering had occurred by chance. Unfortunately no inferential statistics were conducted on this to confirm if this above chance clustering was significant. This finding is extremely robust and was replicated numerous times during that period (Bousfield et al., 1954; Cofer et al., 1966; Hudson, 1968).

More recent studies of semantic clustering have attempted to quantify the degree of semantic relatedness between consecutive items recalled. This is achieved by calculating the Conditional response probability (CRP) of two consecutively recalled items as a function of their semantic relatedness, giving an objective measure of the influence of semantic context on search processes. Steyvers et al. (2004) devised a method for objectively measuring the degree of semantic relatedness of any two items in the English language known as Word Association Space (WAS). This measure assigns a value of 0 to 1 to any word pair; 0 indicating no relation and any value between 0.4 and 1 being a strong association. Using WAS as a measure of semantic relatedness, the semantic CRP indicates that participants tend to recall consecutively words that are more semantically related, at greater than chance level (Kahana et al., 2008).

The most relevant form of context to the present thesis is source. This can be

defined as the origin or properties of a memory. Examples of an item's origin may include its location, study modality or list membership. Properties of an item encompass features such as colour and size of the font a word is printed in (Johnson et al., 1993).

One of the more extensively researched source contexts in terms of memory search is modality. It has been observed that serial-position curves vary as a function of modality, for example whether an item was presented visually or auditorily (Murdock & Walker, 1969). In this study there was a significant advantage for auditory presentation in the recency portion of the serial-position curve, but not during the asymptote. This superiority for auditory presentation was larger at faster presentation rates. The authors concluded that there are two separate short-term memory stores for incoming auditory and visual information; both with differing storage capacities. A larger storage capacity for the auditory store would also account for the auditory superiority in later serial positions.

There is evidence for a notable interaction between modality and temporal distinctiveness of items within a list. Temporal distinctiveness theories of serial position effects assert that recency items in a list are better remembered as they are more temporally distinct than items in earlier list positions (Glenberg, 1987). Furthermore, recall appears to be superior for the auditory modality compared with visual presentation early in the recall period. These tend to be recency items which are more distinct. One explanation is that the auditory modality benefits from more detailed temporal representations of list position than does the visual modality. This produces a more pronounced recency effect for auditory items (Glenberg & Swanson, 1986).

If modality effects can be attributed to separate short-term memory stores for visual and auditory information alone, then there should be no within-modality effects for example clustering according to voice gender or typeface. Hintzman et al. (1972) reported three experiments examining both across and within modality effects using a recognised empirical clustering measure, the Hudson and Dunn index (Hudson & Dunn, 1969). In Experiment 1 participants were presented with mixed-lists of auditory and visually presented words. Experiments 2 and 3 looked for within-modality effects, using two forms of visual (block and script letters) and two forms of auditory presentation (male and female voices). In addition to significantly above chance across-modality clustering in Experiment 1, significantly above chance within-modality clustering was observed in Experiments 2 and 3. Although this alone is not sufficient evidence to categorically refute a theory based on separate visual and auditory short term stores, it does imply a greater level of complexity. Perhaps more convincing is that source recognition judgments for these experiments suggest that participants could distinguish between modalities many minutes after the initial presentation, well beyond the scope of short-term memory. The authors hypothesised that these clustering effects reflect a principle of similarities and differences between items.

Further evidence for a principle of similarity was found by Nilsson (1974). This study demonstrated significantly greater source clustering across modalities (auditory vs visual) than within modalities (male vs female voice, and upper case vs lower case letters). However, participants were able to assign each stimulus to its correct source equally well in all source manipulation conditions. This is consistent with the idea that clustering is a similarity based phenomenon, as two different within modality sources are more similar than two separate modalities sources. Therefore one would expect

higher clustering scores in a list of items where the two sources are less similar. This principle thus far has not been studied for constrained search, i.e. the ability to recall only items from one source while excluding another (e.g recalling only the items spoken by a male voice, and not by the female voice). Significantly greater clustering in dissimilar sources versus similar sources would therefore indicate that participants should be better at constraining search when sources are less similar. This principle of similarity will be explored in Chapter 3.

Frost (1971) presented participants with line drawings in one of four orientations, and found significantly above chance clustering by orientation. There was no significant category clustering for the verbal equivalents of these stimuli, demonstrating that organisation by orientation was not due to semantic associations between pictures within those orientations. It would seem that various types of visual information regarding a pictorial stimulus are accessed during recall and used to guide memory search. This is further evidence that such clustering effects reflect similarities among items.

The role of environmental context on recall is well established (Smith & Vela, 2001). However, fewer studies have investigated search by study environment in unrelated lists. One such study required participants to learn four lists of words in either one, two or four study rooms, before recalling all items in a separate recall room. A very strong effect of number of rooms was observed for clustering by list membership. The mean clustering scores for four study rooms was almost 1.5 times higher than that for one study room (Smith, 1982).

Miller et al. (2013) attempted to find search by spatial location in a virtual town, and therefore a more real world setting. In this experiment participants learned the location of various stores in the virtual town. During subsequent test sessions they

travelled to randomly selected stores, and upon arrival were presented with an item related to that store. This procedure was repeated multiple times over many sessions. After each delivery participants were asked to recall as many of the delivered items as they could. At the end of each session there were final-free recalls for the stores in the town and for all items delivered in that session. Significant spatial clustering was observed both in immediate and both final-free recalls. Temporal associations between items studied proximally in time were found in the recall output, however these did not correlate with spatial clustering. This indicates that the spatial clustering effect was not being driven by temporal associations. The authors postulated that the spatial location in which an item is experienced forms a type of context for that item, even if such context did not exist for the item prior to the experiment. In this experiment when an object is delivered to a store, it is 'flavoured' with the context of the store. Upon recalling an object its context is also evoked, and subsequently cues retrieval of unrelated items that were studied proximally in space.

Research has shown that encoding tasks can be a form of context strong enough to elicit clustering effects. Polyn et al. (2009b) presented participants with an unrelated list of items presented one at a time, each accompanied by one of two concurrent tasks. Halfway through the list the task switched. It was found that items preceding the switch demonstrated a markedly reduced probability of recall. This was accompanied by an increased probability of recall for items following the switch. Also, significantly above chance probability of same-task transitions was found in the task-switching group. The authors believed that the effect was due to the task switch acting as a 'disruptive cognitive event', which caused a shift in context. This in turn isolated a set of items following the switch from those studied prior, causing them to cluster together. It should be noted however that although this study demonstrates a shift in

context as evidenced by differences in recall probabilities near the switch, there was only a single task switch occurring half-way through the list. Therefore there is a strong possibility that task clustering may simply reflect the use of inter-item temporal associations by participants at retrieval.

A second study which partially addressed this issue was conducted by Polyn et al. (2009a). The only notable difference with this study is that task was switched every two to six items, as opposed to once half way through the list. Order of same task trains was also randomised. Significant clustering by task was again observed. An attempt was also made to control for the possibility that inter-item temporal associations may be masking task clustering effects. Using a relabelling technique, lists in a control condition where there was no task switching were randomly assigned the shifting order of one of the experimental lists to create a baseline. This could then be used to calculate the number of expected same-task transitions for each task switch list, if organisation was driven by inter-item temporal associations alone. When averaged across all lists the source clustering effect was large compared with this baseline, suggesting that temporal associations could not completely account for the findings. However a design such as this may still cause an underestimation of source clustering, as same task items were presented in trains. A clearer picture of the strength of the effect may be obtained if task switching was completely randomised. Despite this, it would seem that task switching is a viable method for inducing contextual shifts which produce clustering.

Temporal context plays a fundamental role in all episodic retrieval. Temporal clustering also known as the temporal contiguity effect, is incorporated into and accounted for in a variety of retrieval models (Lohnas et al., 2015; Polyn et al., 2009a; Raaijmakers & Shiffrin, 1981). This can be seen in lists of semantically unrelated items

where items that are studied consecutively tend to be recalled in close temporal proximity. This is typically indexed by calculating CRPs as a function of lag between two consecutively recalled items' respective list positions (lag CRPs). For example in a list comprising five items: Car, Bear, Guitar, Spoon, Arrow; the lag between Car and Bear is +1, Guitar and Arrow is +2 and Arrow and Bear is -3. The probability of each lag in the recall output is calculated to assess the magnitude of the effect. The temporal contiguity effect can be summarised as shorter lags being more common than longer lags (lag recency effect), and forward (+) lags being more likely than reverse (-) lags (Kahana, 1996).

Lohnas and Kahana (2014) argued that the temporal contiguity effect is cumulative. From a meta-analysis of free-recall studies, they found that the temporal contiguity effect was greater when the two previous items in a list were recalled consecutively. This compound cueing effect was interpreted as evidence for a retrieved context account, whereby an item's cue for recall is a recency weighted sum of previous temporal contextual states (Howard & Kahana, 2002; Polyn et al., 2009a). An alternative explanation for this effect is based on rehearsal. It is possible that a contiguity effect would also arise simply because participants use rehearsal strategies based on the serial order of items in a list. The implication here is that temporal information is not encoded, and plays no role in output order. If this account is true one would expect differences in this compound cueing between immediate-free recall (IFR), delayed-free recall (DFR) and continuous-distractor-free recall (CDFR). In CDFR a compound cueing effect should be absent as a distractor task between items should hinder rehearsal, thus severely reducing the temporal contiguity and compound cueing effect. Inconsistent with this account Lohnas and Kahana (2014) found that the temporal contiguity effect and compound cueing effect was present across IFR, DFR

and CDFR.

A related argument for memory not being organised by temporal information was expressed by Hintzman (2016). He argued that all episodic memory tasks have a prospective component. If a participant is aware of the nature of the memory task, they will devise strategies for encoding which will facilitate retrieval. These strategies are often based on serial order of items, leading to a strong lag recency effect in output order. Therefore lag CRPs are a confounded measurement of temporal contiguity, and any experiments that attempt to find evidence for temporal organisation in memory should be those where participants have no knowledge as to the nature of the task.

Spacing-judgment tasks are a good test of this argument. Hintzman, et al. (1975) presented participants with a long list of unrelated items for a later memory test. They were then asked unexpectedly to judge the spacing of two items in the list. The spacings ranged from a lag of one to a lag of twenty-six. It was found that participants were no better at discriminating a lag of one than any greater lag. Furthermore when actual spacings were plotted against mean spacing judgments, the slope was near 0. This indicates that participants did not have the necessary temporal information to make the spacing judgment. Therefore this finding is difficult to reconcile with organisation of memory by temporal information.

Whether the temporal contiguity effect represents memory organised by time or not, it still has important implications for this thesis. In almost all experiments detailed in the following chapters, participants will be asked to recall half of all items presented. If correct items (targets) and incorrect items (source intrusions) are randomly arranged within the same list, then rehearsal strategies or temporal contiguity will hinder recall of targets. In situations where correct and incorrect items

are separated into two separate lists, temporal contiguity will aid recall of targets provided that the retrieval cue for the correct list can be found. This concept will be explored in depth in Chapters 2-4.

Another prominent example of the importance of temporal context in memory is serial-position effects (Murdock, 1962). These early findings describe a serial-position curve whereby items at the beginning and end of lists have a higher recall probability than mid-list items. These are known as primacy and recency effects respectively. In addition, items presented at the end of a list have a significantly earlier mean output position than early and mid-list items. The recency effect in particular has been researched extensively. The recency effect is abolished in DFR tasks, whereby there is a delay between presentation of the last item and the recall period (Glanzer & Cunitz, 1966; Postman & Phillips, 1965). However the effect persists in CDFR, where a distractor task is presented between each item (Bjork & Whitten, 1974; Glenberg et al., 1983; Tzeng, 1973). This is known as the long-term recency effect. This demonstrates that the presence or absence of a recency effect is reliant upon the relative magnitudes of the inter-stimulus interval and the retention interval, as opposed to the absolute magnitude of these delays. The present research will employ DFR tasks to avoid recency items receiving more support than others, thus interfering with constrained search.

1.2.2 - Models and frameworks of memory accuracy control

One of the most prominent attempts to describe memory accuracy control is the Source Monitoring Framework (Johnson et al. 1993). This framework argues that we do not retrieve a source identifier in conjunction with a memory, but that sources (origins) of information are attributed to memories on the basis of evaluative decision processes performed on the memory trace when remembering. Characteristics of

memories that are particularly involved in source monitoring judgments include perceptual information for instance the colour of an item, temporal or spatial context, emotional state or responses at the time of the event, semantic knowledge, and cognitive operations which the individual may have performed on the memory when it was encoded such as elaboration. Source monitoring decisions based on for instance the amount of perceptual detail are often rapid and automatic and do not require conscious awareness of any decision processes, thus the source is identified during remembering. Conversely, source monitoring judgments which require reasoning such as those based on a match with previous semantic knowledge or schema are slower, more deliberate, and may require retrieval of and comparisons with supporting memories.

Source can be identified to varying degrees. For example it is possible to identify who told you a particular fact based on where, when and how they told you, or in the complete absence of where, when and how. Presumably though, the former would be easier given the additional source information available. This is a principle which is yet to be tested for constrained search. It may be possible that if more source information is made available to a participant they may be able to make use of additional source cues to retrieve correct information, making constrained search more accurate. This concept will be explored in detail in Chapter 3. On the whole the Source Monitoring Framework asserts that the accuracy of source judgments is based on the type and amount of source information contained within the memory trace, the similarity between candidate sources (accuracy will be poorer when attributing a memory to one of two male voices, versus a male and a female voice), the effectiveness of the judgment processes used to attribute source, and the criterion the individual uses to distinguish between sources.

An alternative model known as the Strategic Regulation of Memory Accuracy Framework was proposed by Koriat and Goldsmith (1996), to explain how memory accuracy can be controlled directly by an individual in response to situational demands, and how this relates to memory quantity. This approach breaks down post-retrieval processes into two phases, monitoring and control. The model proposes that a 'best candidate answer' to an input query is formed from the combined processes of retrieval and monitoring. This carries with it an assessed probability of being correct, P_a. If the best candidate answer is a total guess, $P_a = 0$. Control processes then compare P_a against a pre-set response criterion P_{rc}. This is derived from the gains of volunteering correct answers versus the costs of incorrect answers. If P_a exceeds P_{rc} the best candidate answer is volunteered, otherwise it is withheld.

It is suggested that recall accuracy in free-report-memory tests are dependent upon three main factors. The first is the effectiveness of the monitoring process. How well does P_a differentiate correct from incorrect answers? The second is how sensitive the control process is. Is the volunteering or withholding of answers sensitive to the output of monitoring? The final factor is what level the participant sets the response criterion P_{rc}.

By allowing monitoring accuracy to vary, this approach allows for a much more complex treatment of the relationship between accuracy and quantity of memory than a simple accuracy/quantity trade off described by signal detection approaches, which rely mainly on criterion placement. For example if a participant has poor memory retention, one would expect very poor accuracy on a forced-choice-memory test. In a free-report test, the same participant's accuracy could still be perfect if their monitoring is excellent, accompanied by a reduction in quantity as they reports only what they remember and monitor as correct. However if a participant's monitoring is

also poor, then it is possible for poor accuracy in a free-report test as well as reduced quantity relative to forced choice.

Both of these approaches have been highly successful at explaining a number of findings in various memory accuracy literatures. Despite this success, both the Source Monitoring Framework and the Strategic Regulation of Memory Accuracy Framework neglect the role of source constrained search in the control of memory accuracy. It is highly unlikely that when asked to recall information, an individual also retrieves vast quantities of related but irrelevant information which they then must monitor as incorrect before the correct answer is found. The focus now centres on models which allow individuals to use source retrieval cues to constrain their search, so that much of this irrelevant information is excluded prior to monitoring.

Koriat et al. (2008), see also Goldsmith (2016), expanded on the original framework to include such a mechanism. The new framework was termed the Metacognitively Guided Retrieval and Report framework (Meta-RAR). Initial pre-retrieval processes use cues in a trial and error fashion, in order to assess the likelihood that sought after information can be accessed in memory. Based on this assessment, a decision is made as to whether to initiate or forego memory search. If a search is initiated, then metacognitive processes establish a search strategy and locate appropriate cues for retrieval.

After an item has been retrieved, post-retrieval processes assess the retrieved information for correctness. If information is judged to be wrong then it is rejected or inhibited. In an advancement on the previous framework, a feedback loop is introduced whereby post-retrieval processes can influence subsequent retrievals or terminate search altogether. If retrieved information is judged not to be sufficiently correct, the search strategy and retrieval cues can be adjusted in an attempt to

retrieve correct information. If the retrieved information is judged to be sufficiently correct, or if it is deemed that finding a better candidate answer is unlikely, then search is terminated.

Finally once search has been terminated, the best candidate answer is assessed for its correctness and a decision is made whether to report it or not. This best candidate answer can be reported with varying degrees of coarseness depending on the individual's confidence in the answer.

This framework provides a key platform for investigating processes such as exclusion of incorrect items from the search set as explored in Chapter 4, and late monitoring processes. However there are two important things to note. The first is that Meta-RAR is only intended to address situations where there is a single answer to a query, for instance "What is the capital city of Costa Rica?" This becomes an issue when attempting to measure how search and monitoring accuracy changes as more items are retrieved, such as from a study list (retrieval dynamics). To make Meta-RAR compatible with the list learning experiments detailed in this thesis, the framework can be modified by allowing retrieval cues to activate a set of items to be searched rather than an individual item. Retrieval then begins using this same cue. Each retrieved item is subsequently monitored for correctness. If correct then the item is monitored as a target, and if it is incorrect then the item is monitored as a source intrusion. Meta-RAR describes a mechanism whereby retrieved information is monitored for whether it is the 'best candidate answer'. A decision is then made whether to continue or terminate search depending on whether the best candidate answer has been found. This is redundant in list learning experiments as there are multiple targets (in this thesis, half of the items in a trial) rather than a single correct

answer. Therefore all references in the model to 'best candidate answer' are disregarded.

Despite these modifications Meta-RAR does not address retrieval mechanisms in any great detail. Given the importance of different forms of context to this thesis and their role in guiding constrained search, it is necessary to examine retrieval models which give a comprehensive overview of the role of context in guiding retrieval and constraining search. One can derive more precise predictions about constrained search if a model is used which explicitly describes how cues/context is used to guide search, and how different forms of context interact during retrieval to influence this process. I shall now review candidate models of retrieval which could fulfil that purpose.

1.2.3 - *Models of retrieval*

One of the most influential models of retrieval to date is Search of Associative Memory, or SAM (Raaijmakers & Shiffrin, 1981). SAM describes two memory stores: A limited capacity short term store (STS) and an unlimited capacity long term store (LTS), acting as an associative network. In SAM the STS is responsible for the rehearsal of information and subsequent transfer to the LTS. Newly studied items initially enter STS. Once the STS has reached full capacity, a randomly chosen item is replaced with a new one. SAM describes that items in STS are always available for recall. In LTS items can be associated with context, other items, or themselves. The longer a pair of items spends together in STS, their strength of their association with each other in LTS increases. Additionally, the strength of a given item's association to context and its self-strength are related to the time spent in STS.

SAM explains that the recency effect in immediate-free recall is due to direct output of available items in STS. The predicted two to five item capacity of STS provides a good fit for the recency effect in serial-position curves. Once there are no

longer items than can be retrieved from STS, the retrieval process continues from LTS. The increase in item associations in LTS is greatest when there are few items in STS such as at the start of a list, thus explaining the primacy effect. This explanation quite adequately explains findings from a variety of studies on primacy and recency effects. Recency effects are also abolished in delayed-recall tasks where the retention interval exceeds the span of STS (Glanzer & Cunitz, 1966; Postman & Phillips, 1965). With regards the contiguity effect, SAM claims that adjacent list items spend longer together in STS, and hence have stronger temporal associations in LTS. Therefore lag-CRPs should be significantly higher for items with contiguous list positions than for longer lags.

However, problems arise for SAM when one considers the well documented phenomenon of long-term recency. Studies have demonstrated that the recency effect persists in continuous-distracter-free-recall tasks, where a participant is asked to perform mental arithmetic of a duration longer than the span of STS during the inter-stimulus interval, and after presentation of the last item (Bjork & Whitten, 1974; Glenberg et al., 1983; Tzeng, 1973).This demonstrates that the presence or absence of a recency effect is reliant upon the relative magnitudes of the inter-stimulus interval and the retention interval, as opposed to the absolute magnitude of these delays. If the recency effect could be explained simply as the output of STS, then a sufficiently large retention interval in a continuous-distractor-free-recall task should abolish the effect but, it seems that this is not the case.

Furthermore, a similar scale invariance exists for the temporal contiguity effect. This effect has been shown to persist in continuous-distractor-free-recall tasks, where subjects undergo sixteen seconds of mental arithmetic during the inter-stimulus interval and during the retention interval (Howard & Kahana, 1999). This makes it

unlikely that the temporal contiguity effect arises from associations formed between items which co-occupy STS. In addition, SAM makes no predictions about forward asymmetry of lag-CRPs in free-recall tasks as observed by Kahana (1996). Finally SAM deals only with temporal contexts. A class of models known as retrieved context models describe the interactions between various forms of context pertinent to this thesis.

One such example is Context Maintenance and Retrieval (CMR) (Polyn et al. 2009a). The model explains that at study, items become associated with a unique combination of two different contextual representations: temporal context and source context, in addition to their pre-existing semantic context. When memory is searched the current state of context is used as the retrieval cue, and the likelihood of any given item being retrieved is driven by the similarity between the current state of context and the contextual features of stored items. The closer the similarity in context, the more likely an item is to be retrieved.

At the start of the recall period the items which are most likely to be retrieved are recency items, as their temporal context will best match the current state of context. The primacy effect is modelled as increased attention to early list items. When an item is retrieved, its source, temporal and semantic contextual features are incorporated into the current state of context, which is then used to guide retrieval of the next item. For instance, if the just retrieved item is from Source A, then it is more likely that the next item retrieved will also be from Source A. However, the next retrieved item is also likely to be one which was studied in a nearby list position due to retrieved temporal context. Recall ends when no more items can be retrieved.

Evidently there is no context reinstatement mechanism at the start of the recall period, so search cannot technically be constrained to correct items. However it would

require only a minor modification whereby the target context can be incorporated in the retrieval cue at the start of the recall period, in order to make the model compatible with constrained search. CMR provides clear and testable predictions regarding how constrained search should progress over a recall period given the interactions of multiple forms of context. The main issue with CMR with regard control of recall accuracy is that it has no monitoring mechanism. Therefore it is restricted to search processes in its utility for investigating control of recall accuracy.

Lohnas et al. (2015) proposed a generate-recognise successor to CMR based on temporal context, termed CMR2. This model was intended to explain effects commonly seen in recall tests of multiple lists, such as proactive and retroactive interference. The model describes that temporal context is not confined to a single list. Rather using a free parameter β_{post}^{recall}, temporal context is allowed to drift across lists during both recall and study phases. Specifically this parameter controls the rate of contextual drift from the recall period of one list to the study period of the next. Allowing context to drift throughout the experiment permits inadvertent retrieval of items from wrong lists.

Like CMR, the cue for retrieval in CMR2 is the current state of context, and the likelihood of retrieval is related to the similarity between an item's context and the current state of context. Therefore β_{post}^{recall} is vital for targeted recall of items from previous lists. The model states that β_{post}^{recall} can be manipulated by a participant depending on the task. If the task is to recall the current list, then the rate of contextual drift between lists will be high to maximise the disparity in context between prior-list items and the current state of context. However if the task is to recall a prior list the participant will reduce the rate of contextual drift between lists, in order to minimise the disparity in context between prior-list items and

the current state of context. Only by doing this can items from previous lists effectively compete for retrieval. In the experiments detailed in Chapter 2 of this thesis which defines source as List membership, participants never have prior knowledge of which list they will be asked to recall. Therefore one would assume that values of β_{post}^{recall} would always be low to enable them to recall either list. Therefore this parameter will not be explored.

The main difference between CMR and CMR2 is a monitoring mechanism, whereby retrieved items are checked for correctness before they are output. This is a fairly simple mechanism. When an item is retrieved, its context is compared to the current state of context and a similarity value, *u* is derived. The other key parameter in the monitoring mechanism is c_{thresh}. In a recall task where the requirement is to recall the current list, items are rejected if *u* does not exceed c_{thresh}.

Recall of a previous list is slightly more complex. CMR2's monitoring mechanism plays a key role in reinstating the target context in this case. At the beginning of the recall period the current state of context is again the retrieval cue. This time the similarity threshold value for *u* takes on a different state, c_{thresh}^{target} . Now items will be rejected if *u* exceeds the threshold value, meaning that target items should not have a strong match with the current state of context. The monitoring mechanism remains in this state until a target-list item is retrieved. When this occurs the threshold value reverts to c_{thresh} so that items with a strong match to context are accepted. This will then allow mostly target items to be retrieved from this point.

Despite the success of CMR2 in explaining many phenomena such as serial-position effects, temporal contiguity and long-term recency in continuous-distractor-free recall, its main drawback is that there is no mechanism for directly reinstating the target context of prior lists at the start of the recall period. This seems counterintuitive

given our ability to retrieve memories of events years in the past (Pillemer et al., 1988). However this is useful for the present treatment, as the notion that such a context reinstatement mechanism does exist can be investigated by testing the predictions of a model where it is absent. Such predictions include significantly poorer accuracy, and significantly slower retrieval for the first recalled item when recalling a prior list compared with recall of the current list. These predictions will be tested in Chapters 2 and 4 respectively.

1.2.4 - *Behavioural studies of constrained recall*

A paradigm that is specifically designed to test the ability to recall subsets of items is the list-before-last paradigm. This paradigm requires participants to recall the list immediately preceding the current list. Target and intervening list lengths are manipulated to provide a measure of ability to isolate the target list. If one assumes a retrieval model where items are randomly sampled from a search set containing mainly correct items, then probability of correct recall should only be affected by the length of the target list. If participants cannot isolate the target list, the intervening list will directly interfere with the memory traces of target items. Longer intervening lists lead to greater interference, which will in turn affect the probability of correct recall of target items. In his seminal study, Shiffrin (1970) found that probability of correct recall was affected by target list length but not intervening list length, indicating that target lists can be isolated. This study also demonstrated that forgetting appears to be a result of retrieval failure rather than trace decay.

Jang and Huber (2008) investigated the processes which drive isolation of the target list in the list-before-last paradigm. They proposed a contextual based account, whereby the participant must reinstate the context of the target list in order to recall it to the exclusion of an intervening list, and that the act of performing list-before-last

recall between the target and intervening list drives a shift in context between the two lists. If there is no list-before-last recall between the target and intervening list this contextual shift does not occur, and isolation of the target list is much more difficult. Participants were presented with a series of lists. Target list length, intervening list length and presence or absence of list-before-last recall were manipulated. It was found that when list-before-last recall was present between the target list and intervening list, only target list length affected correct recall, replicating Shiffrin (1970). When the intervening list directly followed the target list without list-before-last recall, correct recall was affected by length of both the target and intervening lists. Accordingly, incorrect recall of the intervening list was only influenced by intervening list length when recall was present between lists. Therefore, it would appear that the presence or absence of list-before-last recall influences the similarity in context between the target and intervening lists, which in turn affects target list isolation.

Unsworth et al. (2012) conducted a more fine grained analysis of list-before-last recall to examine how effectively participants can isolate the target list. In a departure from the methods of Shiffrin (1970) and Jang and Huber (2008) participants were required to recall either the list-before-last or the current (control) list. The standard finding of an effect of target list length but not intervening list length on proportion recalled was replicated on list-before-last trials. Increasing list length on control trials also reduced proportion recalled as expected. The most interesting finding from this study was that proportion recalled was poorer on list-before-last trials than control trials irrespective of target list length. Therefore the very presence of an intervening list was sufficient to adversely affect performance. In addition analysis of intrusions revealed that on list-before-last trials, intrusions originate

not just from the intervening list, but from lists prior to the target list. Taken together these results suggest that in order to isolate a target list, participants rely on noisy contextual cues to delineate a search set which contains both target and non-target items.

Another interesting example of source-constrained retrieval has been explored in the paradigm testing unconscious plagiarism (aka cryptomnesia). In this paradigm participants initially generate items as a group but must later recall their own ideas – i.e. they must engage in source-constrained recall. Brown and Murphy (1989) developed the first experimental approach. In their study four participants took turns generating exemplars of a semantic category. Each generated exemplar had to be unique to all four participants. This procedure was repeated four times so that sixteen unique exemplars in total were generated, four per participant. Later the participants were given a response sheet and were asked to recall the four exemplars they generated (recall-own), and then to subsequently generate 4 completely new exemplars for that category (recall-new). In the recall-own task 75% of participants recalled at least one exemplar generated by someone else. In total 7.3% of all responses were plagiarised, which was significantly above chance. In addition 29% of participants recalled at least one novel intrusion (item not generated during generation phase). These accounted for 2.3% of all responses. Taken together the evidence demonstrates that source constrained retrieval was imperfect.

Landau and Marsh (1997) explored the role of source monitoring in unconscious plagiarism errors. According to the Source monitoring framework (Johnson et al. 1993), sources which are highly similar should be more difficult to differentiate than sources which are less similar, which will lead to poorer monitoring for sources whose features more strongly overlap. In two experiments Landau and

Marsh explored this principle of similarity. It was predicted that more plagiarism errors would occur in a recall-own task when own and partner ideas were more similar than less similar.

In Experiment 1, participants took turns with a computer to generate solutions to a Boggle word puzzle. For every one solution the participant generated, the computer would generate three. In total sixteen solutions were generated, four by the participant, twelve by the computer. In one condition (reveal) the computer's solutions were revealed to the participant one letter at a time, and the participant needed to guess the solution. In another condition (intact-read), participants were presented with each of the computer's solutions in their entirety. The reveal condition was intended to render own ideas and computer generated ideas more confusable. The cognitive operations involved in searching the puzzle for the computer's solutions are similar to those used to search for one of their own solutions. The cognitive operations required to read the intact solutions are less similar. Therefore there should be less plagiarism errors in the intact-read condition. As predicted, significantly more plagiarism errors were committed in the reveal condition than the intact-read condition. These results suggest that making participants guess the computer's solutions made them more similar to the participants' own responses. This caused a greater number of plagiarism errors to occur.

Experiment 2 implemented a manipulation that was predicted to reduce the number of plagiarism errors, by making the partner/computer solutions more perceptually and contextually distinct from own solutions. During the initial generation of items half of the participants played Boggle with a computer, and the other half played with another person. The procedure for the computer-partner condition was similar to the intact-read condition from Experiment 1. It was predicted that

participants who played with another person would commit fewer plagiarism errors in the recall-own task than those who played with a computer. This is due to the responses being offered by a human partner possessing far richer perceptual and contextual cues, which can be used to distinguish between own solutions and partner solutions, than those offered by a computer monitor. As expected, participants committed more plagiarism errors when they had initially played the game with a computer as opposed to another person. This demonstrated that playing with a human partner improved source monitoring by rendering the partner source more distinctive.

One major issue surrounding these earlier studies of cryptomnesia is that they are not fully counterbalanced. Participants are never asked to recall their partners' ideas, so the relative prevalence of source errors whereby people 'give their own ideas away' is never established. It was never known whether these errors were as common or more common than plagiarism errors. A second issue is that cryptomnesia is described purely as failure of source monitoring. The contribution of constrained-search processes to the avoidance of plagiarism errors was never investigated. This is also the case with much of the multiple-list literature i.e. list-before-last.

Hollins et al. (2016) aimed to solve both of these issues by first including a recall-partner condition, whereby the participant was required to recall their partner's ideas. Secondly to address the issue of the role of constrained search in cryptomnesia, this study employed a variant of free recall known as Externalised-Free Recall (EFR). There were two main instructions. The first was to recall only a subset of items, in this case own ideas or partner's ideas (not both), but also to report any incorrect information that happens to come to mind (wrong source ideas or novel ideas). The second instruction was to write task-compliant retrievals on one side of a response

sheet and non-task-compliant retrievals on the other side of the response sheet. This provides a presumably accurate account of exactly what the participants searched, and monitoring responses for each retrieval. New measures that Hollins et al. were able to access were the total number of task-compliant (target) and non-task-compliant (source intrusion) items generated prior to monitoring. Then, a measure of monitoring accuracy for targets and source intrusions separately can be derived.

Interestingly there was a greater propensity for participants to 'give away' their own ideas than plagiarise those of their partner. This can be traced to greater availability and poorer monitoring of wrong-source ideas in the recall-partner task than the recall-own task. This study demonstrates the critical importance of both fully-counterbalanced designs and investigations of item generation in further cryptomnesia research. The next section will review the EFR literature and detail the form of this paradigm that will be adopted in this thesis.

1.3 - Methods

1.3.1 – *Externalised-Free Recall*

Although standard-free-recall paradigms have contributed much to our understanding of memory search mechanisms, they do have one serious limitation in that they are unable to capture the majority of errors that occur during the search process. It is widely accepted that retrieval occurs in two phases. Initially an item is generated on the basis of its match with a retrieval cue, for example 'current list'. The item is then subject to a highly efficient monitoring process whereby it is assessed for correctness (Watkins & Gardiner, 1979). If the item is deemed to be correct by the monitoring process it is overtly recalled, and if it is deemed to be incorrect then the item is withheld. In a standard-free-recall paradigm, most errors occurring during memory search do not get reported as they are edited out prior to overt recall. The

only errors which are reported are those whereby monitoring has failed. Therefore to effectively measure constrained search a paradigm is needed whereby participants are also required to report incorrect information that comes to mind. This will yield a presumably accurate account of what the participant has searched.

The first attempt at such a paradigm was made by Bousfield and Rosner (1970). Participants in this study were given either one of two recall instructions in a multi-trial-free-recall task. Standard-free-recall instructions were to recall all items they could remember in any order. Instructions for uninhibited-free recall were to report everything that came to mind regardless of its nature during the recall period, even if the participant knew that they were making errors.

Uninhibited free recall instructions resulted in significantly greater numbers of novel intrusions (did not appear on any list) than standard free recall, although these were still rare. Variability in novel intrusion errors was much greater for uninhibited recall than standard recall. To correct for this, a subgroup of participants from the uninhibited condition matched with standard instructions for variability in novel intrusions was taken. Analyses of novel intrusions in the uninhibited-instructions subgroup indicated that there was no difference in frequency of novel intrusions between instructions. The most common category of errors were intra-trial repetitions. Uninhibited-free recall resulted in significantly greater numbers of repetitions than standard-free recall.

Across the first four of five trials there was no significant difference in correct recall between the two instructions. However, correct recall was significantly greater for uninhibited recall on the final trial compared with the final trial of standard recall. Post-hoc analyses revealed that this correct recall advantage can be attributed to increased probability of recalling items from previous lists. This was also shown to be

completely independent of the number of errors (novel or repetitions) reported. It was suggested by the authors that standard-free recall imposes strong inhibition on errors, which also generalises to items that were accessible in previous recall attempts.

A more advanced version of this was presented by Kahana et al. (2005). In this study, young and older adults studied multiple lists of words with a recall test after each list. Participants were told to recall all items from the current list. However, they were also instructed to report all words that came to mind while attempting to recall the current list. In an advancement on the Bousfield and Rosner (1970) procedure, participants were also asked to press a key immediately after reporting a word which they believed was not on the current list.

Both younger and older adults reported large numbers of intrusions per list. However a significant difference in age group was only found for correctly identified intrusions. Older adults were significantly worse at identifying intrusions (previous-list or novel items) than younger adults. Although the main focus of this study was on monitoring deficits in older adults, this demonstrates that EFR is a useful way of measuring errors in generation too owing to the large number of reported intrusions. I shall now review two key studies which used EFR to equally assess both generation and monitoring errors, using analysis methods which will be employed in this thesis.

One of the most useful aspects of EFR for investigating constrained search, is that one can examine the patterns or dynamics of correct and error responses over a recall period. This allows one to investigate for instance at what point constrained search breaks down, and provides a more detailed insight into the relationship between constrained search and monitoring, and their combined contributions to recall accuracy. Unsworth et al. (2010) examined the dynamics of erroneous and correct responses in EFR and the potential theoretical information that can be gleaned

using this procedure. Six lists of ten words were studied, with EFR instructions as described by Kahana et al. (2005) after presentation of each list. Participants were required to recall specifically items from the current list, but to report any other items which came to mind. This study particularly focused upon three types of recall error: Prior-List Intrusions (PLI), Extra-List Intrusions (ELI) and repetitions. PLIs refer to the recall of items presented prior to the target list, ELIs are erroneously recalled words which were never presented at study and repetitions are items recalled more than once within the same list.

Firstly in almost all cases recall would start with a correct response, normally from serial position 1, and then participants would recall five or six correct items consecutively. After output position 7, the proportion of correct responses fell dramatically. These initial correct responses generally came from the primacy end of the serial-position curve. This is unsurprising given that the paradigm employed was a variant of a delayed-recall procedure which generally eliminates recency effects as discussed earlier (Glanzer & Cunitz, 1966; Postman & Phillips, 1965). Conversely, proportions of PLIs and ELIs rose as a function of output position and plateaued around output positions 5 to 10, before gradually falling thereafter. PLIs were significantly more likely to be items presented on the immediately preceding list. This is expected given that items from neighbouring lists will share more temporal contextual features than items from lists with a larger list lag, and are therefore more likely to be recalled consecutively. Finally, PLIs generally clustered in trains of four to five items, roughly half of these originating from the same prior list.

Another aspect of memory search that Unsworth et al. (2010) focussed on was response type at search termination. Participants were significantly more likely to terminate their recall following an error. This is consistent with other studies of recall

termination (Harbison et al., 2013; Miller et al., 2012). When termination data were corrected for frequency of each response type, it was found that participants were mostly likely to terminate their recall following a repetition, a finding later supported by Miller et al. (2012).

In general, Unsworth et al. (2010) observed that participants' monitoring processes were fairly successful. They identified 98% of their correct recalls as correct, 81% of PLIs as incorrect, and 68% of ELIs as incorrect. However, only 47% of repetitions were correctly rejected. When examining monitoring performance as a function of output position, it was found that the probability of correctly rejecting a PLI or ELI increased steadily with output position. When collapsing across output positions, the probability of rejecting an ELI or PLI was significantly higher in the second half of a list than the first. Taken together these results suggested that error monitoring is much more effective later in the recall process. Worse error monitoring in early output positions may be caused by confusion due to greater temporal, semantic or phonological contextual similarity with correct items. An alternative explanation could concern bias in monitoring. In early output positions participants may believe that it is much more likely that they will recall a target than an intrusion. By making an assumption that they have recalled a target, they have a strong tendency to monitor as such without much consideration. Conversely, at the end of the recall period not many potential targets remain. Therefore, an assumption is made that the majority of recalled items will be intrusions, leading to a tendency to monitor as such, again with little consideration.

Another interesting observation was that PLIs originating from more recent lists

were rejected significantly less often than larger lag PLIs. This reinforces the notion that errors that have greater temporal contextual overlap with target items are less likely to be rejected.

The authors interpreted these findings in the context of generate-recognise models of memory adapted to incorporate errors, for example SAM and CMR2. It was suggested that mid-way through the recall process, the list-context cue and previous target-list items no longer serve as effective cues for focussing memory search on the target list. This now weakened cue-target relationship allows for errors which share semantic or temporal contextual features with target items to compete for sampling. As recall progresses correct items get weaker, and more intrusions are generated. While this is a possibility, one cannot confirm this assertion without controlling for the base rate of targets falling as the recall period progresses. Intuitively, constrained search will become more challenging as the number of targets yet to be retrieved reduces. Without separating cue strength from target base rates, it is not possible to conclude that cue strength weakens as the recall period progresses. I will attempt to do this using a computational modelling approach in Chapter 5.

If a participant generates a PLI, it is highly probable that this originated from the list immediately prior to the target list, as the recall contexts of those two lists should be similar. This PLI then serves as a cue for the next item, generally another PLI from the same list as they share temporal contextual features, leading to clusters as observed. ELIs initially occur due to some semantic association with a target-list item. They become clustered as the initial ELI becomes a cue for other erroneous semantic associates. Errors occurring early in the recall process were rejected less often. When these errors were preceded by a correct item, rejection probabilities were lower than

if the error was preceded by another of the same type. This suggests a strong temporal, semantic or phonological contextual overlap between errors recalled early, and correct items. When errors share contextual features with other errors, there is less confusion and rejection probabilities are higher. As the vast majority of repetitions occurred at the last output position, it is likely that the participants knew that they had been recalled before, and chose that moment to end their search.

Unsworth et al. (2013) attempted to specifically examine whether participants can constrain search to a correct list using EFR. Participants completed a number of experimental trials comprising two lists, and control trials comprising a single list. On experimental trials, participants were required to study both lists and then recall one of the two with EFR instructions. On the control trials participants studied just a single list and then recalled that list with EFR instructions.

Participants generated a greater proportion of correct items in the control lists than in either of the experimental lists as expected. Further analyses showed that there was no significant difference in proportion correct between recall of Lists 1 and 2. Participants also emitted a significantly greater number of intrusions in the experimental trials than in control trials. There was again no significant difference in intrusions emitted between Lists 1 and 2.

Analysis of output dynamics however revealed potential differences between List 1 and List 2. For recall of List 1 participants tended to generate correct items early on, but by output position 3 they were no more likely to generate correct items than intrusions. For List 2 participants again searched mostly targets early on. However, by output position 5, they were mostly searching intrusions. In fact for List 1, generation of correct items and intrusions was roughly equal across output positions except for positions 1 and 2. For list 2 correct items are most common early on, whereas

intrusions are more likely than correct items later on. This suggests a difference between the two lists in availability of correct items as the recall period progresses. The proportion of correct responses at output position 1 was also higher for List 2 than List 1, indicating that reinstating List 1 context to initiate recall was challenging.

For monitoring, participants identified significantly more intrusions correctly in the control list than either of the experimental lists. Again there was no significant difference in intrusion monitoring between recall of Lists 1 and 2. On the whole, the main advantage of EFR is the richness of the data that can be gathered and the many different analyses that can be conducted. This allows for a much more in-depth characterisation of constrained search and monitoring than can be achieved with standard-free recall.

Despite the clear advantage of EFR over standard free recall with respect to assessing memory accuracy, the methodology used by Kahana et al. (2005) and Unsworth et al. (2013) is not optimal for assessing monitoring accuracy. Prospective memory research shows that participants frequently forget to make keypresses. In Experiment 1 of Einstein and McDaniel (1990), participants performed a short-term-memory task. At three random points during the task the word ‘rake’ appeared on the computer screen, and participants were required to press a key when this occurred. It was found that young adults were only able to accomplish this on 47% of trials without an external memory aid.

In a memory aid condition participants were allowed to use thirty seconds prior to the task to generate some form of prospective memory aid (facilitated by stationery positioned in front of the participant). Young adults in this condition remembered to press the key on 83% of trials. This was not a case of participants either remembering to press on all three trials or forgetting to press on all three trials, as 29% of young

adults remembered to press on either one or two trials but not all three. In addition, a subsequent questionnaire revealed that participants did think about the prospective memory instruction during the experiment, suggesting that forgetting to press the key was not due to forgetting the instruction, rather a genuine failure in prospective memory. From these findings it is easy to see how monitoring performance in EFR can become conflated with prospective memory errors, where participants forget to reject intrusions using a keypress.

A way to solve this would be to force participants to make a monitoring judgment on each item. Hollins, Lange, Berry and Dennis (2016) successfully achieved this by requiring participants to write task-compliant and non-task-compliant responses in separate columns on a response sheet. The main drawback of this method was that the order of retrievals was not recorded. Therefore analysis of output dynamics as reported by Unsworth and colleagues was not possible.

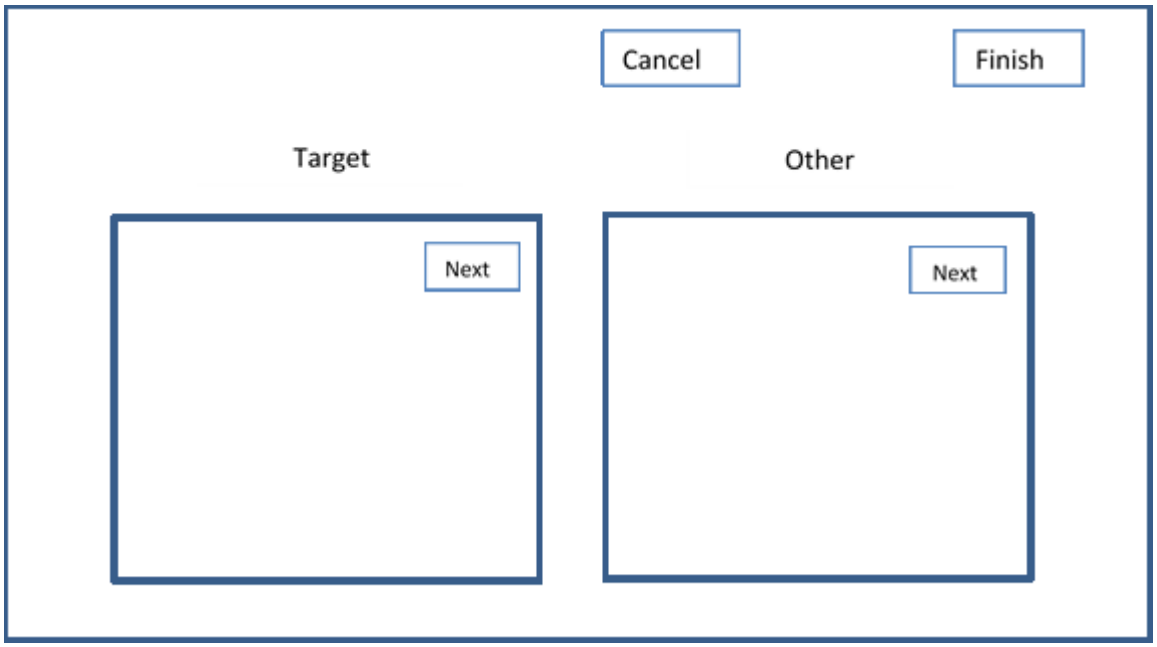
For the EFR studies conducted in this thesis a modification of the Hollins, Lange, Berry and Dennis (2016) method will be employed, whereby responses will be recorded electronically on a tablet device as depicted in Figure 1. Participants will be instructed to recall items from only one of two sources. All items recalled from the correct source should be written in the 'Target' box. However, if any wrong source items happen to come to mind, they should be written in the 'Other' box.

After each reported item the participant will be required to press next, which will clear the screen for the next recall. By doing this response order is recorded, allowing for analysis of output dynamics. The 'Cancel' button should be pressed if the participant believes that they have written an item in the wrong box. Participants press finish when they cannot remember any more items. This method has a clear advantage over older versions of EFR in that ambiguity over measures of intrusion monitoring is

eliminated, while preserving the ability to record output order for analysis of retrieval and monitoring dynamics.

Figure 1.1

Layout of the Tablet Screen for all EFR Experiments in this Thesis



Despite this, one must bear in mind the potential for selective reporting. It is possible that participants may deliberately withhold some responses in an attempt to artificially inflate their constrained search accuracy scores. An indicator of this would be if constrained search was invariant as a function of context type. In this thesis I will be exploring a number of contexts which should affect the accuracy of constrained search considerably. I aim to demonstrate using the new procedure that constrained search accuracy does not simply reflect monitoring accuracy, and that it is sufficiently sensitive to detect predictable differences in constrained search accuracy when they arise i.e. across contexts. In order to verify this methodology, I also contrast the findings of EFR with replication studies using a different methodology based on recall latency analysis, where selective reporting is not an issue.

1.3.2 - *Recall latency analysis*

Modelling of recall latencies or the precise timings of retrievals in recall, can offer a useful measure of the size of a participant's search set. In this approach, recall is modelled as sampling with replacement from a search set which is defined by the retrieval cue. This search set is not necessarily confined to experimentally defined targets, and may include some incorrect information i.e. source intrusions. Items are assumed to be sampled at a constant rate. Therefore, if there are many source intrusions in that search set, it will take longer to retrieve novel targets than if the search set is small and containing mainly targets. Ultimately the mean latency to recall, or average time taken to recall any item reflects the size of the search set (Wixted & Rohrer, 1993). Given that this methodology relies purely on latencies rather than accuracy measures, and does not require the participant to report incorrect information, selective reporting is absent.

An exponentially-modified Gaussian (ex-Gaussian) distribution, which has been shown to be a good fit for response time data in a number of situations (Heathcote et al., 1991) will then be fitted to the data. From this, three parameter estimates can be derived which correspond to different stages of the recall process. Tau (τ) corresponds to the mean time of the exponential phase and is the index of search set size. Mu (μ) and sigma (σ) are the mean and standard deviation of the Gaussian phase. These represent the mean and standard deviation of the delay between initiation of search and recall of the first item. In essence, the time taken to locate the correct retrieval cue. A more detailed description of procedures and analysis methods are presented in Chapter 4. If participants can exclude incorrect items from the search set, estimates of τ should be significantly lower for instructions to recall one source than to recall both sources. Estimates of μ and σ will allow testing of predictions from models such as

CMR2 regarding the initiation of search, for instance whether we are able to successfully reinstate context for recall of a previous list.

Initial attempts by Wixted and Rohrer (1993) to utilise this analysis method to measure search set size were focused on the build-up and release from proactive interference in the Brown-Peterson paradigm. Typically, in this paradigm participants complete four trials. Each trial comprises a rapidly presented list of three words followed by a roughly thirty second numerical distractor task to avoid rehearsal, then a time limited recall period. All of the list items from the first three trials are derived from the same semantic category. Items on the final trial are derived from a different semantic category from the previous nine presented on trials 1-3.

Typically the percentage of items correctly recalled per trial declines progressively over the first three trials. This is known as build-up of proactive interference. On the final trial when the semantic category is swapped, a large recovery in percentage of correctly recalled items is observed. This is termed release from proactive interference. An explanation for this is that as more items from the same category are studied, participants become increasingly unable to distinguish between items presented on the current trial and those presented on previous trials, due to interference among items. Effectively, participants are unable to confine their search to the current trial, and in fact search many items of the same category that were presented since the beginning of the experiment. On the final trial there is no interference from previous trials as the items are from a different semantic category, therefore participants are better able to confine their search to the current trial.

In a modification of the standard Brown-Peterson paradigm as described, Wixted and Rohrer (1993) did not present the change category trial, as release from

proactive interference was not of interest. On each trial before the list was presented, a cue appeared to indicate the semantic category that the list items belonged to.

As predicted, the percentage of correctly recalled items declined progressively with each trial. When the ex-Gaussian distribution was fit, estimates of τ for trials 1-3 were 2.42, 4.21 and 4.40 respectively. This demonstrates a large increase in search set size from trials 1-2 and a much smaller increase from trials 2-3. Estimates of μ and σ did not differ as a function of trial, therefore proactive interference does not influence the onset of recall, in essence the ability to find a retrieval cue. From these findings it was determined that proactive interference in Brown-Peterson trials represents a progressive broadening of the search set, as participants cannot distinguish items from current and previous trials.

It would seem therefore that analysis of recall latencies offers a promising approach to testing set size constriction in source-constrained retrieval. Estimates of τ are sensitive to predictable differences in search set size. μ and σ also appear to map onto early processes described in Meta-RAR such as setting a search strategy and locating retrieval cues.

One of the very few attempts to utilise recall latencies to index focussed memory search was conducted by Unsworth et al. (2013). In the first experiment participants studied trials of either two lists or a single control list. On the two-list trials participants were required to recall one of the two lists as instructed. On the control trials, participants recalled the single presented list. Exponential-cumulative-recall curves were fitted to the data. Although this study does not use the same form of curve as Wixted and Rohrer (1993) it can be assumed that both curves measure essentially the same process. A slower rate of approach to asymptotic recall (λ) is indicative of a larger search set. Estimates of λ were smaller on the two-list trials than

on the control trial, indicating a faster approach to asymptotic recall in control trials. To support this, mean recall latency was longer for the two-list trials than control trials. This suggests that when participants attempt to focus their search on a single list when in the presence of another, they include a number of wrong list items in their search set.

A further experiment was conducted to examine the extent to which participants can focus their search on a single list of items. The paradigm was largely identical to that of the first experiment, except a further condition was added where participants were presented with two lists and asked to recall both of them. Estimates of λ were highest for the control trials, lowest for both lists, and intermediate for List 1 or List 2. This indicates that participants can selectively search for a single list when in the presence of another; however, the search set contains some but not all of the incorrect list items.

A final experiment aimed to see if list distinctiveness improved participants' ability to focus search on a single list. The paradigm was similar to the first experiment; however, on the two-list trials, the two lists comprised category exemplars exclusive to that list. For instance, List 1 could be the names of kitchen utensils, and List 2 could be names of animals. Estimates of λ for recall of List 1 or List 2 were no different to recall of control trials, indicating that semantic distinctiveness improved participants' ability to focus their search.

The experiments in Chapter 4 will be replications of select experiments from Chapters 2 and 3. Participants will study three trials comprising two sources; for instance, two lists. For each trial they will be asked to recall either one of the two sources, or both sources, as indicated by the computer screen. Participants will then speak aloud into a Dictaphone all of the words they can remember from the source/s

indicated by the computer screen. A significant reduction in tau for recall of one source relative to both sources is indicative of successful search set size reduction, and by implication constrained search.

The first experiment will look to replicate EFR findings from Chapter 2, relating to the ability to constrain search to one of two lists. Following this, two experiments will explore whether participants can reduce the size of their search set when attempting to recall a single source, in a mixed-list of two sources. The second of these experiments will investigate the role of source similarity on search set size in Mixed-lists. If EFR does not suffer from selective reporting confounds then the conclusions drawn from these independent replications and their EFR counterparts should be the same. However, this is dependent on a good model fit to the data; otherwise, parameter estimates can be misleading, a concept explored more in Chapters 4 and 5. As such goodness of fit tests will be conducted on the model fits and the behavioural data from these latency experiments examined, to check that the obtained parameter estimates truly reflect participant behaviour.

Using a combination of EFR, computational modelling approaches and recall latency analysis, this thesis will be able to give a comprehensive account of the main stages in control of recall accuracy: Constraining the search set, constrained search and source monitoring. These methods will allow me to test predictions of extant accounts and develop new ideas as to the true nature of memory accuracy control. I will start by thoroughly testing and scrutinising the new EFR paradigm in a study similar to that of Unsworth et al. (2013).

Chapter 2: Source as List membership

2.1 - Introduction

When we remember past events, we are required to mentally reinstate the context of the event and use this as a retrieval cue. One potential context is temporal information, which is evident in the list-before-last paradigm (Shiffrin, 1970). This paradigm requires participants to recall the list before the one most recently presented. The length of the intervening list and to be remembered list is manipulated to investigate the effects of interference and other factors on ability to retrieve items from the target list.

Jang and Huber (2008) manipulated target list length, intervening list length and the activities participants engaged in between lists. Participants studied twenty-four lists of words and were informed that they may be tested for list-before-last recall after any given list. For the lists that list-before-last recall was not tested, participants engaged in some other activity after the study period, such as an n-back or letter completion. They found that only target list length affected recall of the target list when participants engaged in list-before-last recall between the lists. However, target list recall was affected by both target list and intervening list length when other activities were engaged between lists. This suggests that list-before-last recall between target and intervening lists protects the target list from interference, demonstrating that the correct information can be temporally selected given the right experimental conditions.

In these list-before-last studies, participants engage in standard-free recall and so the true number of errors being generated (as opposed to reported) is masked by

monitoring. Participants may have been generating just as many intervening list items as target list items, only for them to be withheld due to not meeting the correctness criterion for monitoring. Unsworth, et al. (2013) addressed this issue with an Externalised-Free-Recall paradigm, as detailed in Chapter 1. Participants were presented with either a single list (control condition) or two lists, the second being presented immediately after the first. In the control condition participants received EFR instructions to recall the single list. In the experimental conditions they were given EFR instructions to recall either List 1 or List 2, which was used to estimate both generation and monitoring for the target list. As expected, participants generated more incorrect list intrusions in both experimental conditions than in the control condition. However, interestingly the number of incorrect list intrusions generated for recall of Lists 1 and 2 were almost identical. This indicates that participants could reinstate the context of List 1 equally as successfully as they could reinstate list 2 context.

However differences were revealed between List 1 and List 2 when output dynamics (the proportion of targets and wrong-list intrusions retrieved as a function of output position) were examined. For recall of List 1 participants begin by recalling predominantly targets, but by output position 3 there was no difference in the reporting of targets and incorrect list intrusions. For recall of List 2 participants again retrieved mostly targets. However, by output position 5 predominantly intrusions were being retrieved. This indicates that there may be differences between the lists in the availability of items during recall.

Another main advantage of EFR is the ability to assess monitoring accuracy of the retrieved items. Unsworth et al. found that participants were significantly better at rejecting intrusions in the control lists than either List 1 or List 2. This was expected

given that there are more types of intrusions that one can retrieve. In control lists, intrusions can only come from prior trials and novel items not presented on any list. For experimental trials intrusions may also arise from another list within the same trial. These wrong-list intrusions are more likely to cause monitoring errors as they are contextually more similar to targets than prior-trial intrusions and novel items, making discrimination more challenging. Interestingly, there was no significant difference in intrusion rejection rates between List 1 and List 2. Taken together, these findings demonstrate that proactive and retroactive interference do lead to intrusion monitoring errors; however, neither to a greater extent. Output dynamics for monitoring were not reported in this study, so it is not known whether there were differences in intrusion monitoring over the course of a recall period. In this thesis, all EFR studies will detail output dynamics of target and intrusion monitoring to gain a fuller assessment of monitoring ability.

The ability to access not only overt retrievals, but those also filtered out by monitoring make EFR an ideal methodology to explore the role of context reinstatement in free recall accuracy. Given that the participant is reporting all incorrect information that comes to mind, one can also investigate the relationship between ability to reinstate context, and the search processes that guide retrieval. Clustering is a potential candidate given that it is a significant predictor of other recall performance measures such as total-free recall, (Brown et al., 1991; Santa et al., 1975; Sederberg et al., 2010) and is a fundamental consequence of the search process (Kahana, 2017).

Clustering is characterised by the grouping together of items at recall which share similar contextual features, such as semantic associations, temporal associations and source features, often yielding a highly organised recall output along the

dimension of similarity among the items (Polyn et al., 2009a). Least is known about source clustering. However, various experiments show clustering effects in Mixed-lists where sources differ along dimensions of modality (Hintzman, et al., 1972; Murdock & Walker, 1969; Nilsson, 1974), visual shape (Frost, 1971), encoding task (Polyn et al. 2009b), spatial location (Miller et al., 2013) and emotional valence (Long et al., 2015). Given that clustering is a consequence of contextually based search, then the magnitude of clustering can be seen as an indicator of the strength of a contextual retrieval cue. Therefore, if constraining search involves reinstatement of context to cue retrieval, then participants who are more efficient at constraining search (reinstating context) should demonstrate superior clustering in their recall output than those who are less efficient at constraining search.

For the experiments in this chapter participants will study items presented in two lists, List 1 and List 2, and be asked to recall just one of the two lists. If we assume a role for context reinstatement in constraining search, and that List 1 is the target list, then participants who are more accurate at constraining search should demonstrate highly clustered recall outputs, with a high proportion of items from the target list; for instance, 1,1,1,1,1,1,2,2. However, those participants who are less efficient at constraining search should demonstrate less clustered recall outputs which contain a larger proportion of non-target-list items; for instance, 1,2,1,1,1,2,2,1,2. If context reinstatement is not involved in constraining search, then there should be no relationship between accuracy of constrained search and magnitude of clustering.

Due to the robustness of clustering effects in standard-free-recall tasks this can also be used to verify EFR as a suitable method to measure the relevant search processes in this thesis. EFR is a significant departure from standard-free recall in that retrieval is uninhibited, and participants are required to pay specific attention to

source. It is important to assess whether these procedural departures interfere with natural search processes occurring with retrieval. Ideally, one would desire clustering and total recall to not significantly differ between standard-free recall and EFR. This will be explored in the present chapter.

A modified version of Meta-RAR (Goldsmith, 2016) provides a useful account of how context reinstatement is used to constrain search. In response to a recall instruction such as 'Recall List 1' (of two), participants choose a retrieval strategy and develop appropriate retrieval cues to recall List 1. Retrieval cues in this instance are analogous to reinstating the target context (List 1). This will hopefully activate mostly List 1 items; however, some List 2 items may be included. Retrieval is then initiated using this same retrieval cue. The main issue is that retrieval mechanisms in Meta-RAR are very poorly defined; therefore, predictions regarding the link between clustering and constraining search cannot be made. For the same reason the model is inappropriate for retrieval dynamics, although one can make predictions regarding overall search and monitoring accuracy across an entire trial.

CMR (Polyn et al. 2009a) on the other hand is a model that attempts to explain retrieval processes in great detail. The model accounts for clustering effects by stating that when an item is sampled at recall, its semantic, temporal and source contextual features are also retrieved. This serves to update the current state of experimental context, which in turn is used as a cue to retrieve the next item. The likelihood of any given item being retrieved is determined by the level of contextual overlap between items in memory and the current state of context. Hence, recall of an item from a particular source is most likely to trigger retrieval of another item with the same source features. At present, CMR does not describe a context reinstatement mechanism; however, this can easily be incorporated by allowing the target source in

addition to the current state of context to define the search set. Ability to reinstate target context could be linked to factors such as how well source has been encoded, which in turn will directly influence the degree of contextual overlap between the target and non-target source.

If a participant can reinstate the context of either list, it is because they have effectively encoded source, and can very easily distinguish between the source contexts for List 1 and 2 i.e. there is a weak overlap in context between the 2 sources. In this case the search set as defined by the reinstated-target-list context will contain few wrong-source items. If retrieval is then initiated using this same contextual cue, it follows that the weak contextual overlap between the two sources will mean that search will remain trained on items from the same source for longer; hence, a high degree of clustering. Therefore participants who are better at constraining search should display stronger clustering in item generation.

There are two scenarios for a total lack of an ability to reinstate target-source context depending on which list is being recalled. In both cases the time of test (current state of) context will define the search set and initiate retrieval. If List 2 is the target list, then the search set will still largely comprise targets given that list 2 items have a much greater overlap with the current state of temporal context than List 1 items. If List 1 is the target list, then no ability to reinstate source context will have a much greater effect on measures of constrained search. If the participant must rely on time of test context to define the search set, then this search set will contain many more wrong-source items and fewer targets than when List 2 is the target list. Retrieval will also most likely start with wrong-source items given the strong contextual overlap between time-of-test (current state of) context and List 2 items.

The main issue with CMR is that it does not describe any formal monitoring

mechanism. Its successor CMR2 (Lohnas et al., 2015) extends the retrieval mechanism outlined in CMR to multiple lists and incorporates a simple monitoring mechanism based on context comparisons, allowing the model to operate as a generate-recognise approach (Watkins & Gardiner, 1979). Instead of each list having its own specific context or tag, context is allowed to drift across lists at encoding. This drift rate can vary to suit the demands of the task. Time-of-test context is the retrieval cue in CMR2, and the likelihood of an item being recalled is influenced by the relative match between current context and all items in memory. For tasks such as list-before-last recall, contextual drift rate across lists needs to be slow to lessen the mismatch between time-of-test context and the target list. While this ability to control drift rate is fundamental to CMR2, it is important to note that in the experiments reported in this thesis, the participants are never aware of which list they will be required to recall. In the present chapter the contextual drift rate across lists will always need to be slow in case participants are asked to recall a prior list; therefore, this feature of the model is largely redundant.

In order to retrieve items from a prior list, the model utilises a context comparison monitoring mechanism. At the start of the recall period time-of-test context cues items which have a strong contextual match. Each retrieved item's context is then compared with the initial retrieval cue. For retrieval of a previous list, an item is accepted if the match in context falls below a threshold value. The retrieved context of this item is then used to cue retrieval of the next. To retrieve the most recent list, an item is accepted if the match between a retrieved item and time-of-test context exceeds a threshold value. For a more in-depth description of the model see Chapter 1, section 1.2.3.

This series of studies had a number of objectives. The primary aim was to test the viability of the modified EFR paradigm for simultaneously measuring constrained search and monitoring. In order to do this, it was important to establish that search processes are largely unaffected by the uninhibited nature of recall and source monitoring instruction in EFR. To this end, recall and clustering by List membership in EFR were compared with those exhibited in a standard-free-recall experiment, and standard-free recall with source monitoring (intended to represent source constrained retrieval). If search processes are unaffected by the nature of EFR, then one would not expect clustering and recall to differ appreciably across the three experiments. However it is not unexpected for there to be greater clustering and recall in source-constrained retrieval and EFR relative to standard-free recall, owing to the increased salience of source. As source is task-relevant in source-constrained retrieval and EFR, participants will pay far greater attention to source in these tasks at encoding than in standard-free recall. More efficient encoding of source may lead to stronger overall item activation, and an increased tendency to cluster by source.

The second aim was to use the modified EFR paradigm to measure whether participants can constrain search to a single list of items in a two-list trial, and then subsequently monitor the output of search successfully. Various predictions were made from CMR and CMR2 regarding the differences between for example recall of List 1 and List 2, which will be detailed later. Meta-RAR will not be considered further in this chapter, owing to its inability to predict clustering effects.

The third aim was to explore the relationship between context reinstatement and search processes in free recall. This gives an indication as to what the role of context is in control of recall accuracy. If context is involved in the regulation of search accuracy, then participants who are better at constraining search should exhibit more

clustering in item generation. If context is primarily involved in source monitoring, clustering should not differ between people who constrain search more effectively and less effectively.

The final aim was to examine the role of encoded temporal context in control of search accuracy. Same source items in List membership experiments are also presented proximally at study, and are separated by time from the other source. In these situations constraining search could be accomplished without using source, simply by activating same source items through their inter-item temporal associations formed at encoding. Two lines of evidence could suggest this. Theories of temporal distinctiveness (Glenberg & Swanson, 1986; Neath & Crowder 1990) state that there is a recall advantage when participants are presented with words in the Auditory modality compared with the Visual modality. This is because serial-order information is supposedly encoded better in the Auditory modality. If participants are using serial-order information/inter-item associations to focus search at least to some degree, constrained search scores and clustering should be superior for the Auditory modality than the Visual modality. To test this, two versions of each experiment were run; one with Auditory presentation and the other with Visual presentation. The second line of evidence for use of temporal context rather than source would be poorer constrained search for recall of List 1 than List 2 as previously stated.

2.2 - Experiment 2.1 (Standard free recall)

Before examining the role of context reinstatement in search processes and constrained search, it was important first to determine if the particular source manipulation employed in this chapter, List membership, was sufficient to elicit significantly above chance clustering effects. Once this can be established, the second aim was to observe if search processes differ as a function of List membership and

Modality. Given that source is task irrelevant in this experiment, the search set will be largely defined and retrieval initiated by temporal context available at time of test. It is possible that this context is not strong enough to activate all List 1 items due to the difference in temporal context between the 2 lists. If this is the case then List 2 recall should be significantly superior to List 1 recall. If participants do have greater access to serial order information or inter-item associations at retrieval in the Auditory modality, then recall and clustering should be superior for the Auditory than the Visual modality.

2.2.1 - *Methods*

2.2.1.1 - *Participants*

Forty University of Plymouth Psychology undergraduates (4 Male, 36 Female, Mean age = 20.50, *SD* = 2.58) participated in this study in return for compulsory course credit in order to pass a module.

2.2.1.2 - *Design*

This experiment contained one within-subjects factor, List membership, and one between-subjects factor, Modality. List membership was defined as whether an item was presented in List 1 or List 2, and Modality was manipulated by Visual or Auditory presentation of items. Participant numbers were allocated a Modality prior to the experimental session by means of random sampling without replacement. Twenty participants were allocated to each Modality. Sixty either visually or auditorily presented words were randomly assigned to one of three, twenty-item trials. For each trial, the items were randomly allocated to one of two lists. Participants completed all three, twenty-item trials over the course of the experiment. Given that the trial procedure was identical for each trial, it was necessary for analyses to include an additional within-subjects factor, Trial number, to see if performance decreased across trials due to fatigue, or increased across trials due to practice. Memory was tested

three times in total; each memory test occurred thirty seconds after the presentation of List 2 in each trial.

2.2.1.3 - *Materials*

Stimuli were sixty semantically-unrelated verbal equivalents of the Snodgrass and Vanderwart (1980) pictures. The original images are line drawings standardised on dimensions of familiarity, visual complexity, name agreement and image agreement. Concepts for the original images (e.g. dog) were selected to be exemplars of common semantic categories; for example, four-legged animals. The Battig and Montague (1969) category norms were used as a guide for compiling the images. All participants received the same words as stimuli.

In the Visual condition, words were presented in the centre of the screen, against a white background; in black Courier New font; size 32. In the Auditory condition stimuli were auditory recordings of the same words as those used in the visual condition, in a real human male voice. The male voice actor spoke each word into a recording device. Audio recordings were made in such a way that each stimulus was represented by a single audio file (.WAV). The computer screen remained blank (white) throughout stimulus presentation. To account for individual differences in hearing ability, volume was adjusted manually to suit the participant prior to the experiment by presenting a series of beeps of different volumes through headphones. Participants were asked to indicate which was the loudest volume that they were comfortable with. Before running any experimental participants, two individuals who were not involved in the study piloted the auditory stimuli to check that they were comprehensible, and represented the intended words. All sixty stimuli were

presented to these pilot participants through headphones. For each stimulus they were asked to write down the word they heard on a sheet of paper. All words reported by the pilot participants matched the respective auditory stimuli.

Stimuli for the practice trials were ten further Snodgrass and Vanderwart (1980) images in their original pictorial form to avoid interference with experimental trial stimuli. These were presented in the centre of the screen, with a height and width of 75% screen size. All participants received the same practice stimuli. None of the verbal equivalents of these stimuli appeared in the experimental trials.

2.2.1.4 - *Procedure*

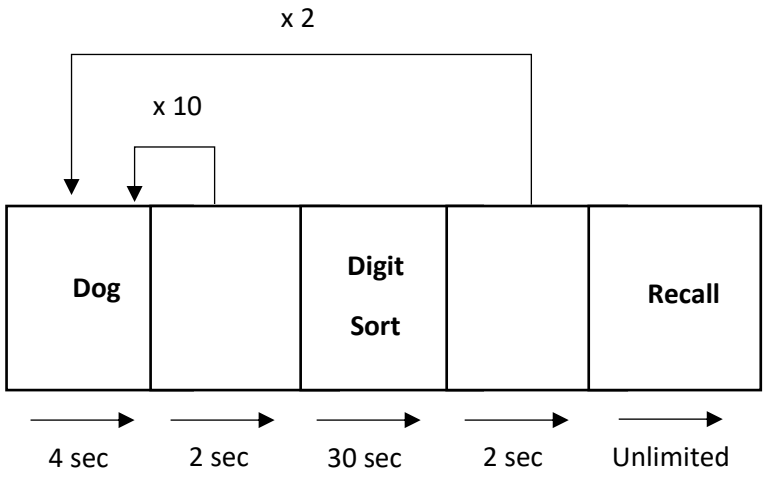
Prior to the experimental lists, the participants completed a single practice list with pictorial stimuli to familiarise them with the procedure. For each experimental trial participants studied twenty words either visually on a computer screen, or through headphones. They were asked to simply remember as many items as they could from both lists (forgetting previous trials) for a later memory test. Each word was presented for four seconds, with a two second Inter-stimulus interval. In the Auditory condition, each spoken word was of different duration; therefore, stimuli were presented over a four-second period, with silence filling time when words were not being spoken. After the tenth item, the participants were presented with three digits in the centre of the computer screen accompanied by the phrase; 'Say these three digits in descending order' (Courier New font; black text; size 32; white background). As instructed, participants read the three digits out loud to the experimenter in descending order. Each set of three digits remained on the screen for three seconds before being immediately replaced by another. Ten sets of three digits were displayed in total. This digit-sorting task served as the dividing line between the two lists. After a

two-second pause, the second list was presented in an identical manner to the first. The participants then repeated the digit-sorting task described earlier with ten different sets of three digits. This prevented the use of short-term memory at recall. The participants were then given a tablet device for the test phase.

For the test phase the computer screen displayed the phrase; 'Write down as many items as you can on the pad'. Participants then pressed a 'start' button on the tablet, which presented them with an on-screen box in which they could write all the words that they could remember, using a stylus. After each word had been written, the participant pressed a 'next' button in the top right corner of the box to clear it for the next item. Previously recalled items were not visible to the participant after pressing 'next'. When the participant could not recall any more items, they pressed a finish button which closed the application. At this stage the tablet was returned to the experimenter. This recall period was not time-limited. On pressing the space bar on the computer keyboard, the prompt; 'Are you ready for the next trial?' appeared on the computer screen. When the participant pressed the space bar again, a blank screen appeared for two seconds, followed immediately by the study phase for the next trial. After the test phase of the third trial had concluded, the participant pressed the space bar and the message; 'The experiment is now over, thank you for your participation' appeared on the screen. See Figure 2.1 for a schematic depiction of the experimental paradigm.

Figure 2.1

Schematic Depiction of a Single Trial of the Experimental Paradigm for Experiment 2.1.



Note. Digit sort = Distractor task where participants were asked to mentally rearrange and speak aloud the three digits on the screen in descending order. There were ten of these over a thirty second period. This task was used for all experiments in this thesis. This diagram depicts the Visual condition only. For the Auditory condition, the computer screen was blank throughout word presentation. Text was only displayed during the digit sorting task and at recall. Recall instructions were to recall as many items as they could.

2.2.1.5 - *Scoring*

2.2.1.5.1 - Recall

Ultimately the aim of these three experiments is to compare search processes across experiments which do not contain the same number of 'correct' items; therefore, the total number of items recalled is difficult to interpret. As such, recall is expressed as a proportion correct score in this chapter. For the present experiment, recall was compared across Lists and Modalities. Given that there are ten items in each list, the proportion correct score for each list (PcRecall) will be the total number of items recalled by the participant in each list, divided by the total number of items in a single list (10), as shown in Equation 2.1.

$$PcRecall = \frac{n_l}{10} \tag{2.1}$$

Where n_l is the number of items recalled by a participant in list number l.

2.2.1.5.2 - Clustering

The unit of clustering employed for analysis was consecutive recall of two items presented in the same list, on a given trial (repetition). The number of repetitions occurring in each participant's recall output was used to calculate a value for Adjusted Ratio of Clustering (ARC), proposed by Roenker et al. (1971). This is essentially a proportion of the number of above chance repetitions observed in the recall output, to the maximum possible number of above chance repetitions for that output. The formula yields scores of 0 for chance clustering, 1 for perfect clustering and <0 for below chance clustering. Each participant's ARC score was the mean of ARC values for all three trials. In the present series of experiments ARC can be expressed as in Equation 2.2:

$$ARC = \frac{R - E(R)}{maxR - E(R)} \tag{2.2}$$

Where R represents the number of observed repetitions. maxR is the maximum possible number of repetitions for a given output and is expressed as in Equation 2.3:

$$maxR = N - k \tag{2.3}$$

Where N is the total number of items recalled by the participant, and k is the number of sources in the trial (two in all cases). E(R) is the expected number of repetitions given chance clustering and is expressed as in Equation 2.4:

$$E(R) = \frac{\sum_i n_i^2}{N} - 1 \quad (2.4)$$

Where n_i is the number of items recalled from category (source) i and N is as before. Trials where the participant recalled fewer than four items were discarded. This is because an ARC score for N = 3 could only be calculated based on a maximum of one repetition (maxR = 1). Therefore very subtle differences between recall outputs yield vastly different ARC scores. In an experiment comprising two sources, a participant who recalled three items with the sources ordered 1,2,1 would score ARC = -2 whereas a recall output ordered 1,1,2 would yield ARC = 1 or perfect clustering. It was decided that all EFR experiments presented in this thesis would adopt this four-item restriction for consistency.

It should be noted that it is theoretically possible for participants to possess clustered 'knowledge' of sources, yet retrieve items in an order that will yield very low ARC scores. For example, a recall output with sources ordered 1,2,1,2,1,2 is highly organised; however, there are no same source repetitions meaning that ARC will be extremely low. Although there is potential for this to occur, it is expected that recall will be driven largely by temporal associations formed between proximally presented items at encoding, as described by CMR. Therefore, recall outputs should resemble something of the form; 1,1,1,2,2,2, yielding a high ARC score.

2.2.2 - *Results*

All Bayesian analyses in this thesis were performed using the BayesFactor package (Morey & Rouder, 2018) in R (R Core Team, 2019). Single sample *t*-tests were conducted to observe if List membership was a sufficient source manipulation to elicit source clustering. Minimum detectable effect size for these *t*-tests with assumed

power of .8 was calculated as d = 0.76. Participants' recall outputs exhibited significantly above chance clustering in both the Visual modality (M = 0.60, SD = 0.34), $t(19) = 7.88$, $p<.001$, $d = 1.76$, $BF_{10} = 4.92 \times 10^4$ and the Auditory modality (M = 0.55, SD = 0.36), $t(19) = 6.76$, $p<.001$, $d = 1.51$, $BF_{10} = 1.29 \times 10^4$. Given the scale of ARC and the large effect size, these could be considered moderate clustering effects. Bayes Factors demonstrate extremely strong evidence for above chance clustering in both Modalities. A 2 (Modality: Auditory, Visual) x 3 (Trial number: 1,2,3) mixed-ANOVA was conducted to investigate whether clustering differed as a function of Modality and across trials. Given the sample size and assumed power of .8, the minimum detectable effect size for the main effects of Modality and Trial number were $\eta_p^2 = .11$ and $\eta_p^2 = .03$ respectively, and $\eta_p^2 = .03$ for the interaction. There was no significant main effect of Modality, $F(1,38) = 0.20$, $p = .66$, $\eta_p^2 = .005$, $BF_{10} = 0.33$ no significant main effect of Trial number, $F(2,76) = 0.58$, $p = .56$, $\eta_p^2 = .02$, $BF_{10} = 0.13$, and no significant interaction, $F(2,76) = 1.90$, $p = .16$, $\eta_p^2 = .05$, $BF_{10} = 0.54$.

These results demonstrate that clustering by List membership did occur in participants' recall outputs for both Modalities; however, clustering did not differ between Modalities. Clustering did not differ as a function of Trial number either, indicating that context based search did not differ throughout the experiment. Finally, Modality and Trial Number did not seem to interact. Sensitivity analyses indicate that the experiment was insufficiently powered to detect an effect of Modality or Trial number. However, Bayes Factors did provide credible evidence that these factors did not affect clustering although the evidence for a lack of interaction was inconclusive.

A 2 (List membership: 1,2) x 2 (Modality: Auditory, Visual) x 3 (Trial number: 1,2,3) mixed-ANOVA was conducted to investigate whether recall differed as a function of List membership, Modality and across trials. Given this sample size and

assumed power of .8, the minimum detectable effect size for List membership and Trial number was η_p^2 =.02, and η_p^2 = .10 for Modality. Required η_p^2 for interactions was .02. There was found to be no significant main effect of List membership, $F(1,38)$ = 1.80, p = .19, η_p^2 = .05, BF_{10} = 0.58, Modality, $F(1,38)$ = 0.01, p = .93, η_p^2 <.001, BF_{10} = 0.31 or Trial number $F(2,76)$ = 0.52, p = .60, η_p^2 = .01, BF_{10} = 0.06. There was also no significant interaction between List membership and Modality, $F(1,38)$ = 0.20, p = .66, η_p^2 = .01, BF_{10} = 0.24, Modality and Trial number, $F(2,76)$ = 0.01, p = .99, η_p^2 = .003, BF_{10} = 0.08, List membership and Trial number, $F(2,76)$ = 1.48, p = .23, η_p^2 = .04, BF_{10} = 0.29, or a three-way interaction, $F(2,76)$ = 0.82, p = .45, η_p^2 = .02, BF_{10} = 0.25. Despite the fact that the experiment is again insufficiently powered to detect an effect of Modality, Bayes factors indicate credible evidence for no effect of Modality. Bayesian analysis also found inconclusive evidence for no main effect of List membership and credible evidence that these factors did not interact in any way. See Table 2.1 for recall in each Modality and List membership condition collapsed across trials.

Table 2.1

Means and Standard Deviations for Proportion of Items Correctly Recalled as a Function of Modality and List Membership Across Procedures.

Procedure	PcRecall Auditory List 1		PcRecall Auditory List 2		PcRecall Visual List 1		PcRecall Visual List 2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
SFR	0.46	0.22	0.51	0.18	0.48	0.23	0.50	0.19
SCR	0.56	0.24	0.58	0.22	0.55	0.26	0.56	0.25

Note. SFR = Standard-Free Recall; SCR = Source-Constrained Retrieval

2.2.3 - *Discussion*

The most important and reassuring finding from the present experiment was that participants do exhibit source clustering by List membership in their recall outputs. This suggests that at least in standard-free-recall experiments, search can be

driven by this source manipulation, which is a foundation for the subsequent source-constrained retrieval and EFR experiments to be compared against. Despite the lack of power to detect an effect of Modality using NHST, it seems from the Bayesian analysis that there is credible evidence that participants did not derive any clustering or recall advantage from the Auditory modality. This suggests that inter-item temporal associations were not the main driving force behind search.

An alternative and highly likely explanation may be related to the nature of the task. Modality effects often manifest as enhanced recency effects for the Auditory modality compared to the Visual modality (Murdock & Walker, 1969). In delayed-free-recall tasks such as the experiments presented in this chapter, recency effects are eliminated (Postman & Phillips, 1965). Therefore, there may not have been an advantage to be gained from the Auditory modality in this instance, at least for search processes. In addition, Neath and Crowder (1990) state that it is specifically serial-order information that is better encoded for the Auditory modality, not temporal information generally. Search may still be driven significantly by inter-item temporal associations, encoding strategies or the temporal contiguity effect. This does not mean to say that serial-order information is not useful for monitoring, which will be explored in the next experiment.

There also appears to be no difference in item availability between the two lists, indicating that participants can just as easily activate items from List 1 with time-of-test context as they can List 2 items. In terms of CMR2, the data imply that the model can only explain this pattern if it assumes the rate of contextual drift across lists is minimal. Finally, item availability and contextual based search did not seem to differ as a function of Trial, thus ruling out fatigue or practice effects. This also indicates that participants were not changing their encoding or retrieval strategies as the experiment

progressed. The next experiment will explore whether participants are able to accurately identify an item's List membership when it is recalled.

2.3 - Experiment 2.2 (Source-constrained retrieval)

This experiment sought to extend the findings of Experiment 2.1 in order to understand the impact of source salience on search processes in free recall. In this experiment, participants were required to recall as many items as they could from each trial, and for each item they recalled they needed to indicate whether the item was presented in List 1 or List 2. The key difference between this task and standard-free recall is the salience of source. Participants were instructed that they would need to provide source monitoring judgments for each recalled item; therefore, they would explicitly encode list membership of each item at study. Given that source is likely to be better encoded in this task than standard-free recall, one would expect source to play a greater role in driving search than in Experiment 2.1, manifesting as a greater degree of clustering. Recall may also improve as overall item strength could benefit from better encoding of source. Given the increased prominence of source in the present experiment, Modality effects on clustering which tend to reflect search driven by temporal context should be small if not absent.

It was not expected that source monitoring should differ as a function of List membership. Regarding Modality effects, it is possible that List membership is not a feature of an item which can be monitored; therefore, monitoring judgments may be accomplished using temporal context and serial-order information. For instance, one may make a List membership judgment based on whether the retrieved item was presented before or after the one that was retrieved before. If this is the case, better encoding of serial order information in the Auditory modality may improve monitoring.

2.3.1 - *Methods*

2.3.1.1 - *Participants*

Forty further Psychology undergraduates from the University of Plymouth were recruited for this study (11 Male, 29 Female, Mean age = 19.83, *SD* = 1.26) in exchange for compulsory course credit in order to pass a module.

2.3.1.2 - *Design*

The design was largely identical to Experiment 2.1, and the same random sampling with replacement method of participant allocation to Modality conditions was applied. As with Experiment 2.1, an additional within-subjects factor, Trial number, was added to the analysis to check for fatigue or practice effects. The only procedural difference between this experiment and Experiment 2.1 was at test, where participants were required to make source monitoring judgments for each recalled item. As before, two versions of the experiment were run, with Visual and Auditory presentation of items to investigate Modality effects on search and monitoring. Memory was again tested three times. Each recall task occurred thirty seconds after presentation of List 2.

2.3.1.3 - *Materials*

Stimuli were identical to those used in Experiment 2.1.

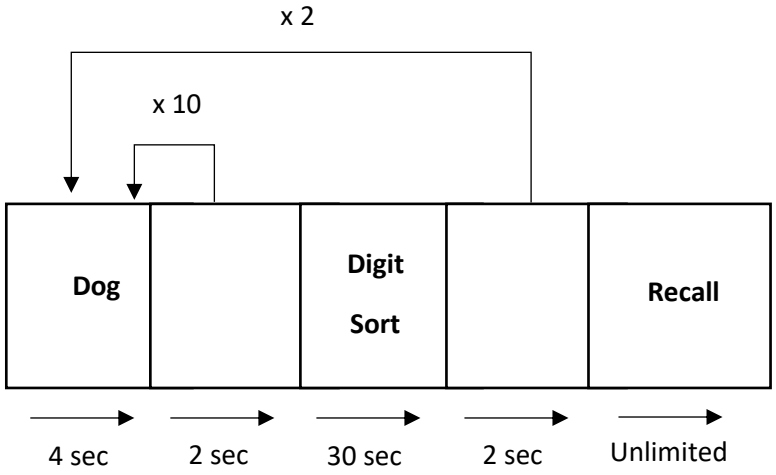
2.3.1.4 - *Procedure*

The study phase was very similar to that of Experiment 2.1. The only difference was that participants were instructed to remember which list each word was presented in (forgetting previous trials), in addition to the words themselves. At the start of the test phase, the computer screen displayed the instruction 'Write down as many items as you can on the pad' in the centre of the screen with the additional instructions 'Left = List 1, Right = List 2' underneath. When participants pressed the

start button on the tablet they were this time presented with two boxes, one on the left side and the other on the right side of the tablet screen. The left hand box had the heading ‘List 1’, and the right hand box had the heading ‘List 2’. For each recalled item, the participants wrote the word in the box corresponding to the list they believed the item was presented in using a stylus. Items presented in List 1 were to be written in the left box, and items that were presented in List 2 were written in the right box. When participants began writing in one of the boxes, the other would disappear in order to prevent writing across both boxes. After writing each word, participants pressed ‘next’ to clear the screen for the next word. Previously recalled words were not visible to participants. If participants believed they had accidentally written an item in the wrong box, they pressed a cancel button at the top of the tablet screen to reset the screen, enabling them to write the word in the correct box. When participants could no longer remember any more items, they pressed a finish button in the top right corner of the screen to close the application. This procedure was repeated twice more for the remaining two trials. Transition between test phase and study phase was identical to Experiment 2.1. See Figure 2.2 for a schematic diagram of the experimental paradigm.

Figure 2.2

Schematic Depiction of a Single Experimental Trial for Experiment 2.2



Note. Digit sort = Digit sorting distractor task used throughout this thesis. This illustration depicts the visual condition only. For the auditory condition, text was only displayed on the screen for the digit sorting task and at recall. Recall instructions were to recall as many items as the participant could remember, writing List 1 items in the left box and List 2 items in the right box of the tablet screen.

2.3.1.5 - *Scoring*

2.3.1.5.1 - *Recall and clustering*

Recall and clustering were scored identically to Experiment 2.1.

2.3.1.5.2 - *Source Monitoring*

To assess a participant's monitoring efficiency, a proportion correct score PcMonitor was calculated. This is the proportion of all items recalled that were attributed to the correct list. This is expressed as in Equation 2.5:

$$PcMonitor = \frac{A_a + B_b}{A_a + A_b + B_a + B_b} \quad (2.5)$$

Where A and B represents the list that the item was presented in, and a and b is the participant's monitoring response. For instance A_a is an item presented in list A that the participant has attributed to list A. This value was averaged across all three trials to

gain a value of PcMonitor for each participant.

2.3.2 - *Results*

Firstly, a 2 (Experiment: 2.1,2.2) x 3 (Trial number: 1,2,3) mixed-ANOVA was conducted on the recall data collapsed across List membership and Modality, to see if the addition of a monitoring instruction significantly influenced recall, and whether there were fatigue or practice effects caused by repetition of the same procedure in each experiment. For assumed power of .8, the minimum detectable effect size for the main effects of Experiment and Trial number were η_p^2 =.06 and η_p^2 =.01 respectively, and η_p^2 = .01 for the interaction. This revealed that the proportion of items correctly recalled did not significantly differ as a function of Experiment, $F(1,78) = 3.53$, $p = .06$, $\eta_p^2 = .04$, $BF_{10} = 1.32$. There was also no significant main effect of Trial number, $F(2,156) = 1.88$, $p = .16$, $\eta_p^2 = .02$, $BF_{10} = 0.23$. Finally there was no significant interaction between Experiment and Trial number, $F(2,156) = 2.80$, $p = .06$, $\eta_p^2 = .03$, $BF_{10} = 0.83$. See Table 2.2 for recall and clustering descriptive statistics across experiments. Bayes Factors suggest that there was only credible evidence for a lack of a main effect of Trial number.

As with Experiment 2.1, a 2 (Modality: Auditory, Visual) x 2 (List membership: 1,2) x 3 (Trial number: 1,2,3) mixed-ANOVA was conducted to observe if recall differed as a function of List membership and Modality within the experiment, and whether performance changed in any way as the experiment progressed. Minimum detectable effect size with .8 assumed power for the main effect of Modality was $\eta_p^2 = .10$, and $\eta_p^2 = .02$ for the main effects of List membership and Trial number. Minimum detectable effect size with .8 assumed power for the interactions was $\eta_p^2 = .02$. There was found to be no significant main effect of Modality, $F(1,38) = 0.08$, $p = .78$, $\eta_p^2 = .002$, $BF_{10} = 0.33$, and no significant main effect of List membership, $F(1,38) = 0.73$, $p = .40$, $\eta_p^2 =$

.02, BF_{10} = 0.22. However, there was a significant main effect of Trial number, $F(2,76)$ = 3.31, p = .04, η_p^2 = .08 although the Bayesian evidence for this was weak, BF_{10} = 1.36. There was no significant interaction between Modality and List membership, $F(1,38)$ = 0.06, p = .81, η_p^2 = .002, BF_{10} = 0.23, no significant interaction between Modality and Trial number, $F(2,76)$ = 0.13, p = .88, η_p^2 = .003, BF_{10} = 0.09, no significant interaction between List membership and Trial number, $F(2,76)$ = 0.38, p = .68, η_p^2 = .01, BF_{10} = 0.10, and no significant three-way interaction, $F(2,76)$ = 1.41, p = .25, η_p^2 = .04, BF_{10} = 0.28. Bayesian evidence fully supports the traditional analyses except for a failure to find sufficient evidence for a main effect of Trial number. Again, it would seem that the experiment was insufficiently powered to detect Modality effects; however, the Bayes Factor does suggest that Modality does not affect recall.

To further explore the significant main effect of Trial number, Bonferroni corrected *t*-tests were conducted to investigate which trials significantly differed when the data were collapsed across Modality and List membership. None of the p-values for the pairwise comparisons reached significance due to the alpha correction indicating insufficient power; however, by examining the PcRecall means for the 3 trials (Trial 1: *M*=0.54, *SD*=0.22; Trial 2: *M*=0.54, *SD* = 0.24; Trial 3: *M*=0.61, *SD*=0.26), it would seem that participants recalled the same proportion of items for Trials 1 and 2, but recalled a greater proportion of items by Trial 3, indicating practice effects. Bayesian pairwise comparisons were not conducted as the evidence for a main effect of Trial number in the omnibus ANOVA was inconclusive. Overall, it seems more likely than not items were equally available in both lists, and that Modality did not affect recall. However, participants did recall a greater proportion of items as the experiment progressed, suggesting practice effects.

Next, a 2 (Experiment: 2.1,2.2) x 3 (Trial number: 1,2,3) mixed-ANOVA was

conducted to observe if the addition of a monitoring instruction significantly affected clustering, and whether clustering changed over the course of the experimental session. Assuming power of .8, the minimum detectable effect size for the main effect of Experiment was $\eta_p^2 = .06$, and $\eta_p^2 = .01$ for the main effect of Trial number. Minimum detectable effect size for interaction was $\eta_p^2 = .01$. There was found to be no significant main effect of Experiment, $F(1,78) = 3.59$, $p = .06$, $\eta_p^2 = .04$, $BF_{10} = 1.00$, no significant main effect of Trial number, $F(2,156) = 2.29$, $p = .11$, $\eta_p^2 = .03$, $BF_{10} = 0.34$, and no significant interaction, $F(2,156) = 1.18$, $p = .31$, $\eta_p^2 = .01$, $BF_{10} = 0.21$. See Table 2.3 for recall and clustering across experimental procedures. It would seem therefore that contrary to predictions, clustering did not increase with the introduction of a monitoring instruction; however, it is likely that this is due to insufficient power in the analysis, also evidenced by an inconclusive Bayes Factor. In addition, there were no detected practice or fatigue effects, which the experiment was sufficiently powered to detect. Unfortunately the Bayesian analysis was only able to find conclusive evidence for a lack of an interaction.

A further 2 (Modality: Auditory, Visual) x 3 (Trial number: 1,2,3) mixed-ANOVA was conducted to investigate whether clustering differed as a function of Modality, and if performance changed across trials. Given an assumed power of .8, the minimum detectable effect sizes for the main effects of Modality and Trial number were $\eta_p^2 = .11$ and $\eta_p^2 = .03$ respectively, and $\eta_p^2 = .03$ for the interaction. There was no significant main effect of Modality, $F(1,38) = 0.11$, $p = .74$, $\eta_p^2 = .003$, $BF_{10} = 0.28$, no significant main effect of Trial number, $F(2,76) = 3.05$, $p = .05$, $\eta_p^2 = .07$ although the Bayes factor was inconclusive, $BF_{10} = 1.01$, and no significant interaction, $F(2,76) = 1.16$, $p = .32$, $\eta_p^2 = .03$, $BF_{10} = 0.31$. Despite a lack of power to detect a significant effect of Modality, Bayesian analyses suggest good evidence for no effect of Modality, or a Modality by

Trial number interaction. However, the evidence for a main effect of Trial number is almost completely inconclusive, given a BF_{10} close to 1. See Table 2.2 for descriptive statistics of clustering by Modality and Trial number.

Table 2.2

Means and Standard Deviations for Clustering Scores in each Modality across Trials.

Score	Auditory						Visual					
	Trial 1		Trial 2		Trial 3		Trial 1		Trial 2		Trial 3	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
ARC	0.60	0.31	0.78	0.45	0.68	0.45	0.59	0.39	0.73	0.49	0.85	0.22

Note. ARC = Adjusted Ratio of Clustering.

Table 2.3

Means and Standard Deviations for Proportion of Correct Items Recalled and Clustering Scores Across Experimental Procedures.

Performance measure	Standard-Free Recall		Source-Constrained Retrieval		Externalised-Free Recall	
	M	*SD*	*M*	*SD*	*M*	*SD*
PcRecall	0.49	0.18	0.56	0.18	0.52	0.20
ARC	0.57	0.35	0.71	0.28	0.64	0.38

Note. ARC = Adjusted Ratio of Clustering. In Externalised-free recall (EFR) participants were only asked to recall one of the two lists in a trial. Therefore, PcRecall for EFR is calculated as the number of items recalled from the correct list divided by 10 rather than 20 for the other procedures.

Single sample *t*-tests were conducted to observe whether participants could monitor the List membership of each recalled item at above chance level. This was defined as a PcMonitor score of 0.5. Assuming power of .8, minimum detectable effect size was $d = 0.58$. Collapsed across List membership, participants were able to monitor List membership when items were studied in the Visual modality ($M = .97$, $SD = .04$), $t(19) = 57.38$, $p<.001$, $d = 26.38$ $BF_{10} = 1.73 \times 10^{25}$, and the Auditory modality ($M = .99$, $SD = .03$), $t(19) = 69.36$, $p<.001$, $d = 31.36$ $BF_{10} = 3.89 \times 10^{26}$. Bayesian analyses revealed

extremely strong evidence for an ability to monitor the List membership of items in both Modalities.

There is little theoretical reason to suspect that there would be an effect of List membership on monitoring; therefore, this was not examined. However, it is perfectly feasible that participants may be using serial order information to assist with monitoring judgments. If this information does form the basis of at least some monitoring judgments then monitoring scores should be higher in the Auditory modality where encoding of serial order information is more effective. A 2 (Modality: Auditory, Visual) x 3 (Trial number: 1,2,3) mixed-ANOVA was conducted to examine if there was a difference in monitoring ability between the Modalities, and whether monitoring performance changed over the course of the experiment. Given an assumed power of .8, the minimum detectable effect size for the main effects of Modality and Trial number were $\eta_p^2 = .11$ and $\eta_p^2 = .03$ respectively, and $\eta_p^2 = .03$ for the interaction. There was found to be no significant main effect of Modality, $F(1,38) = 2.02$, $p = .16$, $\eta_p^2 = .05$, although the Bayes Factor was inconclusive, $BF_{10} = 0.45$. A lack of a main effect is likely due to insufficient power in the analysis. However, there was a significant main effect of Trial number, $F(2,76) = 4.10$, $p = .02$, $\eta_p^2 = .10$, supported by a Bayes Factor, $BF_{10} = 3.18$. There was no significant interaction between Modality and Trial number, $F(2,76) = 1.20$, $p = .31$, $\eta_p^2 = .03$, although the Bayes Factor was inconclusive, 0.34.

To explore the main effect of Trial number further, the data were collapsed across Modalities, and Bonferroni corrected *t*-tests (alpha = .02, minimum detectable effect size for .8 assumed power was 0.53) were conducted to investigate which trials differed in monitoring accuracy. There was no significant difference between Trial 1 (M=.96, SD=.08) and Trial 2 (M=.99, SD=.05), $t(39) = 1.64$, $p = .11$, $d = 0.34$, or Trial 2

and Trial 3 (*M*=1.00, *SD*=.02), *t*(39) = 1.31, *p* = .20, *d* = 0.30. However, there was a significant difference between Trial 1 and Trial 3, *t*(39) = 2.57, *p* = .01, *d* = 0.58. It may be the case that the inability to find an effect for Trials 1 and 2 and Trials 2 and 3 was due to the reduction in power caused by the Bonferroni correction, given the respective effect sizes of the analyses being small to moderate. For Bayesian pairwise comparisons, see Table 2.4. Bayesian post-hoc analyses in this thesis were conducted using a method proposed by van den Bergh et al. (2020), see Appendix A for details. Posterior odds show credible evidence for no difference between Trial 1 and Trial 2, and no difference between Trial 2 and Trial 3. However, there is insufficient evidence to make conclusions regarding the comparison between Trial 1 and Trial 3.

Table 2.4

Bayesian Post-Hoc Analyses for Main Effect of Trial Number on Source Monitoring in Experiment 2.2.

Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
Trial 1	Trial 2	0.22	0.58	0.13
Trial 1	Trial 3	0.22	3.05	0.69
Trial 2	Trial 3	0.22	0.38	0.08

2.3.3 - *Discussion*

This experiment was designed firstly to examine the impact of source salience on search processes in free recall. Unfortunately, conclusions regarding recall and clustering across Experiments 2.1 and 2.2 are difficult to draw, due to insufficient power and inconclusive Bayesian evidence for an effect, or lack of an effect, of Experimental procedure on recall or clustering.

It would seem likely that recall is not affected by List membership. In CMR2 terms, this means that like Experiment 2.1, temporal contextual drift across lists at encoding would need to be slow enough for there to be a sufficient overlap between

the context of List 1 items, and time of test context in order that they can be accessed with the same ease as List 2. There was very little evidence also to suggest that Modality had any effect on recall or clustering. This implies that additional serial order information provided by the Auditory modality did not affect search. However, one cannot discount the fact that effects of Modality on search may have been cancelled out by the fact that this experiment was delayed recall rather than immediate recall, as previously discussed.

The second aim of this study concerned whether participants could monitor the List membership of each recalled item. Results showed that participants were exceptionally proficient at this task, with monitoring scores being near ceiling for each Modality. Unfortunately it was not possible to draw conclusions about the effects of Modality on source monitoring, probably due to a combination of low power and ceiling effects.

One interesting finding in the present experiment was apparent practice effects for recall and source monitoring. In the present experiment participants were specifically asked to attend to the List membership of each item. It seems that participants became more proficient at this with each trial. This was accompanied by an increase in recall between Trials 2 and 3. Although this was not statistically paralleled by a similar increase in clustering across trials, by examining the visual condition in Table 2.2 there is a clear monotonic enhancement in clustering as the experiment progresses. Combined, these practice effects suggest improved encoding of source context as the experiment progressed. Of course, participants may have improved in their ability to monitor source across trials, however this explanation alone could not account for the practice effects on recall and clustering. An

explanation based on encoding of source context would also account for why clustering did not improve across trials in Experiment 2.1. Clustering is indicative of the strength of a contextual retrieval cue; hence, how well context has been encoded. As participants in Experiment 2.1 were not prompted to attend to the context (List membership) of each item, merely the item itself, there is no reason to suspect that encoding of context and therefore clustering would increase with practice. The final study of this chapter will examine whether participants are able to constrain their search to a single list of words, and monitor the output of that search.

2.4 – Experiment 2.3 (Externalised-Free Recall)

This experiment had a number of aims. The first was to test the new modified Externalised-free-recall (EFR) procedure as a viable method of simultaneously investigating the contributions of source-constrained search and source monitoring to the control of recall accuracy. For this to be the case it was important to ascertain that the unorthodox recall instructions of EFR did not interfere with contextually based search processes. Therefore, the indices of search common to all three experiments, overall recall and clustering were compared. Given that the instructions are to constrain search to a single source, but to write down any incorrect items that come to mind, source is obviously highly salient in this task. Therefore, one would expect recall and clustering to be at least equivalent to source-constrained retrieval (Experiment 2.2) and superior to standard-free recall without a monitoring instruction (Experiment 2.1). If the instructions of EFR do interfere with normal search processes, then one would expect to see significantly reduced recall and clustering relative to the other two experiments.

The second aim was to investigate whether participants can successfully constrain their search to one of two presented lists. If this is the case, they should be

able to preferentially retrieve items from the target list. As a reference point, predictions are made using a model which does not possess a formal context reinstatement mechanism, CMR2. This model makes the assumption that all retrieval is accomplished using the current state of experimental context. This is perfectly adequate to recall List 2 as participants can use inter-item temporal associations to retrieve targets, and then terminate search when they cannot remember any more from the target list. Selectively retrieving List 1 is more challenging when there is no context reinstatement mechanism. Time-of-test context will likely activate multiple items from the incorrect list, leading to more wrong list items being included in the search set. It is also likely that participants will retrieve many wrong-list items before retrieving the first item from the target list. Therefore, if there is no context reinstatement mechanism for recalling prior lists as CMR2 suggests, then constrained search accuracy should be significantly poorer for recall of List 1 than List 2 aggregated across all items output. This difference in accuracy between the two lists should be particularly pronounced for the first few items output. If participants can successfully retrieve a prior list, then there should be no difference in constrained search accuracy between the two lists either across the whole recall period or at individual output positions.

The third aim was to illuminate the role of context in the control of memory accuracy, by examining the relationship between constrained search and clustering. As previously stated, clustering is a consequence of contextually based search. If participants attempt to reinstate the target context in order to search for correct items, then constrained search accuracy should correlate positively with the degree of clustering in recall outputs. If there is no relationship between constrained search accuracy and clustering, then context is unlikely to play a role in constraining search.

The fourth aim was to examine to what extent participants can monitor the output of their memory search. Despite there being arguably no perceptual source information associated with List membership, as demonstrated by Experiment 2.2, participants can still make accurate source monitoring attributions based on decision making or strategic processes i.e. feasibility judgments, as explained by the Source Monitoring Framework (Johnson et al. 1993). Such judgments are slower and more deliberate than those based on perceptual information regarding stimulus features. Retrieval begins rapidly in free recall and slows exponentially over time (Wixted & Rohrer, 1993), so for the first retrieval participants may not have the time to assimilate the necessary source identifying information to make a source monitoring judgment, meaning that source intrusion monitoring accuracy should be low. Following this, in addition to the natural exponential slowing of retrieval, the method of item-by-item reporting used in this thesis artificially slows retrieval as participants may be generating items faster than they can report them, affording participants more time to make a monitoring judgment on each item. Therefore, by output position 2 there should be a substantial improvement in source intrusion monitoring accuracy, followed by a more gradual increase thereafter in accordance with natural exponential slowing of retrieval.

The final aim was again to examine the role of incidentally encoded temporal context in constraining search. If participants are largely using temporal or serial-order information to constrain search, then one should observe higher constrained search accuracy scores in the Auditory modality than the Visual modality, due to better encoding of such information offered by the Auditory modality.

2.4.1 - *Methods*

2.4.1.1 - *Participants*

Forty-eight members of the general public were recruited for this study (16 Male, 32 Female, Mean age = 23.17, *SD* = 4.52). They were paid £4 for half an hour of study time.

2.4.1.2 - *Design*

The design of this study was very similar to Experiments 2.1 and 2.2. The only difference was in the recall instructions for the test phase detailed in section 2.4.1.4. Trial number was also removed as a factor in analyses as participants would be asked to recall different items on each trial, and clustering, search accuracy and monitoring accuracy may be dependent on recall instruction. Again, memory was tested three times over the course of the experimental session. On each trial recall was implemented thirty seconds after presentation of List 2.

2.4.1.3 - *Materials*

Stimuli were identical to those used in Experiments 2.1 and 2.2.

2.4.1.4 - *Procedure*

The study phase was identical to that of Experiment 2.2. Participants were instructed to remember as many words as they could from the two lists (forgetting previous trials); in addition to, which list each word was presented in. In the test phase, the participants were presented with the message 'Choose: A or B' in the centre of the computer screen. They were informed that A and B corresponded to one of the two lists that they had just studied, chosen at random. The participant then pressed one of these keys, and the statement 'Recall List 1' or 'Recall List 2' appeared on the screen. The order of these statements across trials for a single participant was counterbalanced by participant number and Modality so that each statement occurred

an equal number of times, for each Modality across the experiment. The statement remained on the screen for the duration of the test phase. At this point, the participant was given the tablet device.

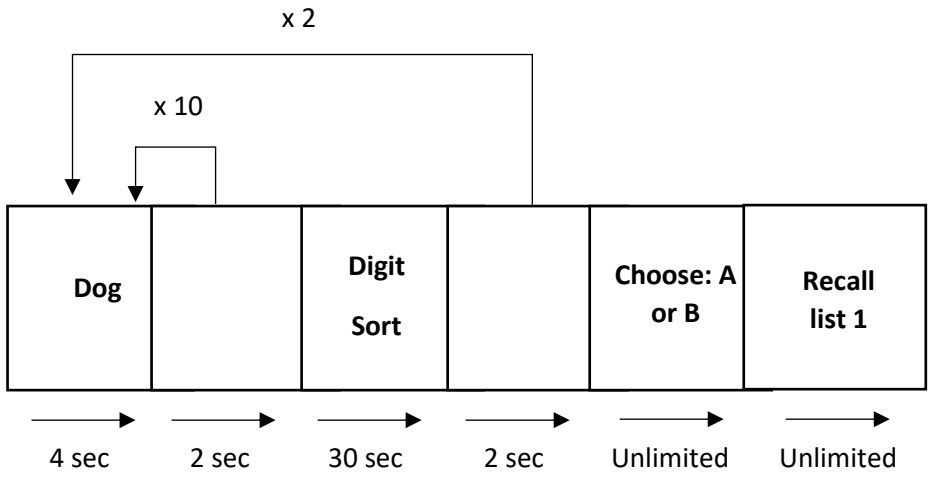
As in Experiment 2.2, when the participant pressed the start button, they were presented with two on-screen boxes. This time the left hand box was labelled 'Target', and the right hand box was labelled 'Other'. The test phase was partitioned into two recall attempts. For the first recall attempt, participants were instructed to recall only the items that were presented in the list indicated on the computer screen, and to write those using a stylus in the 'Target' box on the tablet. However, if any items happened to come to mind from the wrong list, they should be written in the box labelled 'Other'. Participants were again instructed to write down items one at a time, pressing the 'next' button in the top right corner of the box after each item, and that if they accidentally wrote an item in the wrong box they should press 'cancel' to reset the screen. Previously recalled items were again not visible after pressing 'next'. Participants were directed to press 'Finish' when they could not remember any more items from the list indicated by the computer screen (see Figure 2.4).

After a four second blank screen, the 'Target' and 'Other' boxes reappeared on the tablet as before. For the second recall attempt, participants were instructed to recall only the items displayed in the opposite list to the one they had attempted to recall during the first recall attempt, and to write these down in the 'Target' box. If any items from the list that they had tried to recall during the first attempt happened to come to mind, they should be written in the 'Other' box. Partitioning the test phase into two recall attempts gave participants the opportunity to selectively recall all items. This allowed comparisons of total correct recall across all three experiments. However on later reflection, this presents a significant confound for calculating

measures of search. It is quite probable that many items written down in the second recall attempt were still 'in mind' from the first recall attempt; therefore, they cannot be counted as new retrievals. As it cannot be determined how often this occurred, the second recall attempt is excluded from all analyses reported henceforth. See Figure 2.3 for a schematic representation of the experimental paradigm.

Figure 2.3

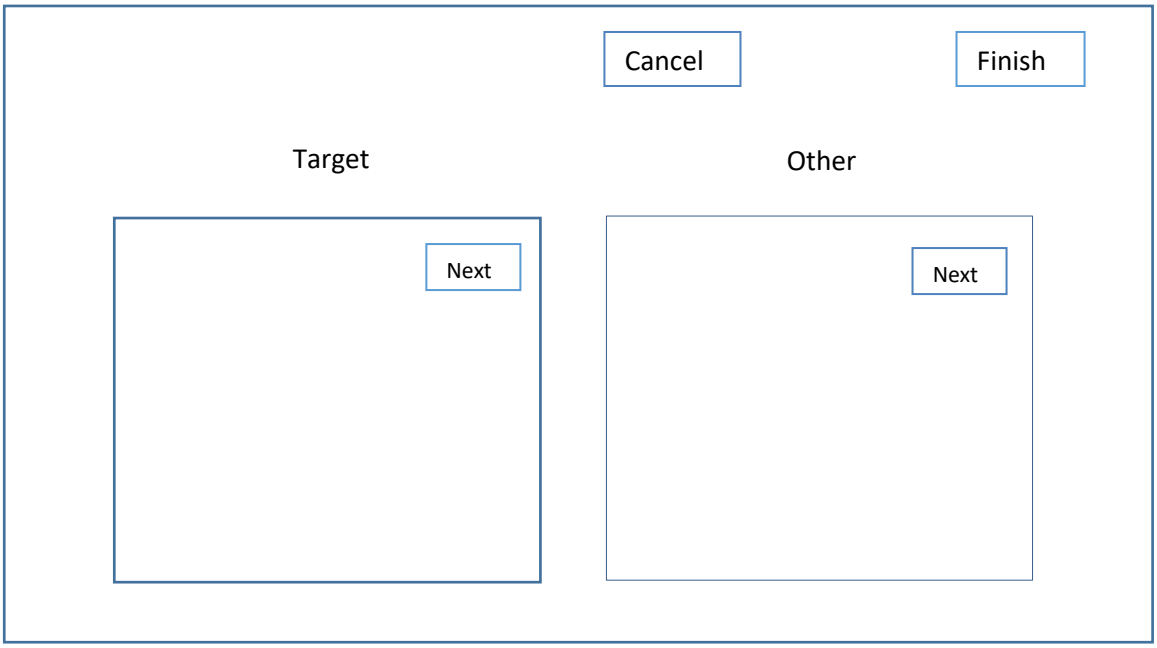
Schematic Representation of a Single Trial for Experiment 2.3



Note. Digit sort = Digit sorting distractor task used throughout this thesis. This diagram only represents the visual condition. In the auditory condition, words were presented through headphones as opposed to the computer screen. For the recall period, the instructions could read either 'Recall list 1' or 'Recall list 2'.

Figure 2.4

Depiction of Tablet Screen as it Appears in Experiment 2.3.



Note. For Experiment 2.2 'Target' and 'Other' were replaced by 'List 1' and 'List 2'. For Experiment 2.1 there was a single box and no cancel button.

2.4.1.5 - *Scoring*

2.4.1.5.1 - *Recall*

As previously stated, comparing recall across the three experiments presented in this chapter is complicated by the fact that participants are required to only recall half of the total number of items in a trial for EFR. In section 2.2.1.5.1, I presented a measure PcRecall to express the proportion of items recalled across modalities and across lists. This will also be used to compare the proportion of items recalled per trial in each experiment. Given that participants are required to recall all items in a trial for Experiments 2.1 and 2.2, PcRecall for these experiments will be calculated by dividing the number of items correctly recalled by the participant by the total number of presented items in the trial (20), as expressed in Equation 2.6.

$$PcRecall = \frac{N}{20} \quad (2.6)$$

Where N is the total number of items recalled in the trial. For EFR PcRecall is calculated as the number of correct items (targets) generated divided by the total number of targets in a trial (10), as expressed in Equation 2.7.

$$PcRecall = \frac{T}{10} \quad (2.7)$$

Where T is the total number of targets generated. Therefore, PcRecall represents the availability of correct items in each experiment.

2.4.1.5.2 - *Clustering*

Clustering was calculated in an identical fashion to the previous two experiments.

2.4.1.5.3 - *Search accuracy*

To assess accuracy of memory search a measure termed PcSource was devised. This is the proportion of the total items correctly recalled that were from the target list. This is expressed as in Equation 2.8:

$$PcSource = \frac{T}{T + S} \quad (2.8)$$

Where T and S represent the number of targets and source intrusions (SI) generated respectively. This value is then averaged across all trials to obtain a mean PcSource score for each participant. PcSource was also calculated for individual lists to observe whether search accuracy is different when participants are asked to recall List 1 versus List 2. For each participant this would be averaged across trials of the same recall instruction to gain a PcSource value for that list. PcSource was also calculated for

each output position across participants to gain a more fine-grained insight into the processes underpinning constrained search. Note that all three trials from each participant were used in the retrieval dynamics analysis (not averaged across trials for individual participants) to maximise power for the analysis. As such, the number of data points contributing to output positions 1-4 (any trials where less than four items were retrieved in total were discarded) was the total number of trials where participants had recalled four items or more, per condition or collapsed across conditions as appropriate.

2.4.1.5.4 - *Source monitoring*

Monitoring performance was partitioned into target monitoring accuracy and source intrusion monitoring accuracy to observe if these behave differently under different manipulations, both across an entire recall period and at individual output positions. Again these will be calculated across trials and across recall instructions (lists). Target monitoring is calculated using Equation 2.9

$$Target\ monitoring = \frac{T_t}{T_t + T_s} \tag{2.9}$$

and source intrusion monitoring is calculated using Equation 2.10

$$Source\ intrusion\ monitoring = \frac{S_s}{S_s + S_t} \tag{2.10}$$

Where T and S are targets and source intrusions, and t and s are the participant's monitoring response. T_t is a target correctly monitored as a target, T_s is a target incorrectly monitored as a source intrusion, S_s is a source intrusion correctly monitored as a source intrusion, and S_t is a source intrusion incorrectly monitored as a target. These formulae were also utilised to calculate a target monitoring accuracy and a source intrusion monitoring accuracy score for each output position. Again, all trials

from each participant were included in the analysis to maximise power in the same manner as for search dynamics.

2.4.2 - *Results*

Three participants were excluded from the analysis as they failed to generate the required four items on any of their three trials. As comparisons are being conducted between experiments whose participants are from different populations (Undergraduates for Experiments 2.1 and 2.2, and the general public for Experiment 2.3), it was necessary to examine whether differences in age between these populations may account for potential differences in recall and clustering across procedures. A one-way between-subjects ANOVA was conducted to compare ages between the three procedures. This did indeed reveal that ages significantly differed, $F(2,125) = 13.66$, $p<.001$, $\eta_p^2 = .18$. Post-hoc Tukey Tests revealed that the age of the general public population (Experiment 2.3) significantly differed from the Undergraduate populations of Experiments 2.1 and 2.2 at $p<.001$. There was no significant difference between the two Undergraduate populations. This was supported by an equivalent Bayesian ANOVA, $BF_{10} = 4.40 \times 10^3$. All corrected Bayesian posterior odds show that the same participant populations differ (see Table 2.5).

Table 2.5

Bayesian Post-Hoc analyses for Age Differences Between Participant Populations

Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
Exp 2.1 (UG)	Exp 2.2 (UG)	0.22	0.60	0.14
Exp 2.1 (UG)	Exp 2.3 (GP)	0.22	22.98	5.16
Exp 2.2 (UG)	Exp 2.3 (GP)	0.22	977.92	219.78

Note. Exp = Experiment, UG = Undergraduate, GP = General Public. Posterior odds are used to identify differences in all Bayesian multiple comparisons. (See Appendix A for multiplicity correction procedure).

Despite there being a statistically significant age difference between the Undergraduate (UG) and General public (GP) populations, the difference in mean ages between the GP and youngest UG population was only 3.34 years. Therefore, it is highly unlikely that age in reality would affect recall and clustering comparisons between the three experiments in any meaningful way.

In order to test whether the recall instructions of EFR significantly affect search processes, two, one-way (Experiment: 2.1,2.2,2.3) between-subjects ANOVAs were run. The first examined recall and the second examined clustering, both as a function of experimental procedure. With assumed power of .8, the minimum detectable effect size for these ANOVAs was η_p^2 =.07. Collapsed across List membership and Modality, the proportion of correct items recalled did not significantly differ between the three experiments, $F(2,122) = 1.65$, $p = .20$, $\eta_p^2 = .03$, $BF_{10} = 0.30$. This was also true of clustering, $F(2,122) = 1.55$, $p = .22$, $\eta_p^2 = .02$, $BF_{10} = 0.28$. Although there seems to be insufficient power to detect significant effects, the Bayes Factors demonstrate credible evidence for a lack of an effect of procedure on correct recall and clustering. Therefore, one can say that the recall instructions of EFR do not appear to affect either correct item availability or context based search. See Table 2.3 for descriptive statistics of recall and clustering across experiments.

The second aim of this experiment was to investigate whether participants could selectively search for items from a target list when in the presence of another. The first indicator of this is whether participants can generate significantly more targets than source intrusions, irrespective of which list was the target list. Two, one-tailed *t*-tests were conducted to test this. For assumed power of .8, the minimum detectable effect size was $d = 0.43$ for these and the following dependent *t*-tests. The first two tests revealed that participants generated significantly more targets than

source intrusions when List 1 was the target list, $t(44) = 6.23$, $p<.001$, $d = 1.48$, $BF_{10} = 1.91 \times 10^5$, and when List 2 was the target list, $t(44) = 6.18$, $p<.001$, $d = 1.56$, $BF_{10} = 1.66 \times 10^5$. Bayes Factors demonstrate extremely strong evidence for successful selective search in both cases. Next, further *t*-tests were conducted to observe if target and source intrusion availability differed between the two lists. If participants can reinstate the target context, then neither should differ as a function of list. A *t*-test revealed that there was no significant difference in target recall between the two lists, $t(44) = 1.44$, $p = .16$, $d = 0.17$. Although this may be considered underpowered, this effect size is small; therefore, target generation may not in fact differ between the two lists. The Bayesian analysis unfortunately cannot shed further light on this as the Bayes Factor was inconclusive, $BF_{10} = 0.42$. A further *t*-test revealed that there was no significant difference in source intrusion availability between the two lists, $t(44) = 0.06$, $p = .95$, $d = 0.01$, supported by a Bayes Factor, $BF_{10} = 0.16$. Although again underpowered, it is highly unlikely that source intrusion generation differed between the lists as the effect size is very small.

An alternative way of looking at this is to examine what proportion of the total number of items generated were targets, (PcSource) and whether this exceeds chance performance (0.5). For the following one-sample *t*-tests, with assumed power of .8 minimum detectable effect size was $d = 0.38$. Collapsed across List membership and Modality, participants could successfully constrain search at above chance level ($M = 0.72$, $SD = 0.19$), $t(44) = 7.72$, $p<.001$, $d = 1.15$, $BF_{10} = 2.23 \times 10^7$. Broken down into individual lists, participants could successfully constrain search when List 1 was the target list ($M = .71$, $SD = .21$), $t(44) = 6.52$, $p<.001$, $d = 0.97$, $BF_{10} = 4.89 \times 10^5$, and when List 2 was the target list, ($M = .72$, $SD = .22$), $t(44) = 6.69$, $p<.001$, $d = 1.00$ $BF_{10} = 8.40 \times 10^5$. Bayes Factors indicated strong evidence for these effects. Importantly, PcSource

did not differ between the two lists, $t(44) = 0.55$, $d = 0.08$, $p = .59$, supported by a Bayes factor, $BF_{10} = 0.19$. Although underpowered it is unlikely that the two lists differed in constrained search accuracy as the effect size was so small, and given conclusive evidence from the Bayesian analysis. See Table 2.6 for targets, source intrusions and PcSource as a function of List membership.

Table 2.6

Number of Targets and Source Intrusions Generated and Overall Search Accuracy as a Function of List Membership.

Search measure	List1		List2	
	M	*SD*	*M*	*SD*
Targets	5.01	2.14	5.38	2.31
SI	2.09	1.78	2.08	1.91
PcSource	**0.71**	0.21	**0.72**	0.22

Note. SI = Source intrusions. Bold text indicates significantly above chance performance. M = Mean, SD = Standard Deviation

So far, it appears as though across an entire recall period, there is no difference in constrained search ability between List 1 and List 2; however, the lists still may differ at various stages in the recall period. Indeed, CMR2 predicts that search accuracy at output position 1 should be poorer for List 1 than List 2 due to a lack of a context reinstatement mechanism. Therefore, search accuracy as a function of output position will be calculated (search dynamics). It should be noted that because not all participants generate the same number of items, analyses lose power as output position increases because the proportion of the original data remaining progressively decreases. As such, with NHST statistics null results become increasingly more likely as the recall period progresses, simply as a function of reducing sample size. Therefore, only Bayesian analyses will be reported as this will indicate whether conclusions should be drawn by quantifying the evidence for or against the null hypothesis.

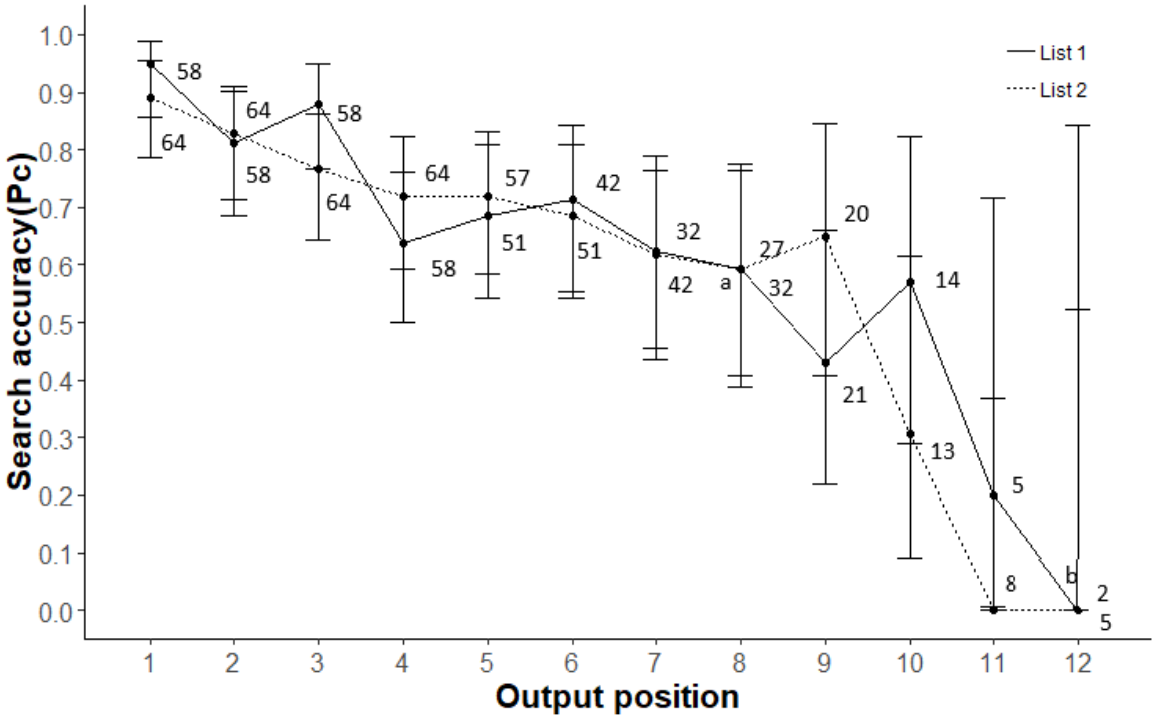
For retrieval dynamics, output positions were collapsed into three bins; early, middle and late, to maximise the potential for finding credible Bayesian evidence for either the alternative or null hypothesis. For instance if twelve output positions are being compared, bin size would be four. This is particularly important for output positions where few participants contribute data (generally late output positions for search and early output positions for source intrusion monitoring). In cases where one condition comprised more output positions than another, output positions from the longer condition that extend beyond the shorter condition were disregarded. This is justified, as in all cases the proportion of total data contributing to these additional output positions in the longer condition was very small. The vast majority of data in both conditions were covered by the same number of output positions.

A series of Bayesian contingency tables were conducted on each bin to detect differences in search accuracy between the lists throughout the recall period. BF_{10} for early (1-4), middle (5-8) and late (9-12) output positions were 0.10, 0.13 and 0.30 respectively. An important hypothesis was that if participants can successfully reinstate a target context, then search accuracy between the lists should not differ at output position 1. BF_{10} for output position 1 was 0.23 indicating no difference between the two lists. From this it can be seen that there is evidence that the lists do not differ in search accuracy throughout the recall period. Crucially, there was no difference between the two lists at output position 1 which indicates successful context reinstatement. Search dynamics for Lists 1 and 2 are depicted graphically in Figure 2.5. The general pattern appears to be very high search accuracy at the beginning of the recall period, followed by a progressive decrease with output position. It was deemed highly unlikely that constrained search would differ as a function of Modality owing to

the fact that this was a delayed free recall task. Therefore, Modality effects were not followed up further.

Figure 2.5

Search Accuracy by Output Position as a Function of List Membership



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below data points indicate the number of trials contributing data to that output position.

[a] List 1 = 27 trials, List2 = 32 trials

[b] List 1 = 2 trials, List 2 = 5 trials

The next point of interest was to illuminate the role of context in constraining search. If context reinstatement plays a key role in retrieving a target list, participants with a greater ability to constrain search should exhibit stronger clustering in their recall outputs, given that clustering is an indicator of successful contextually based search (Polyn et al. 2009a). To test this, participants were divided into High search and Low search conditions according to their PcSource scores averaged across trials, collapsed across List membership and Modality. The High search condition comprised

the participants with the twenty-three highest PcSource scores, and the Low search condition comprised the participants with the lowest twenty-two PcSource scores. Given assumed power of .8, the minimum detectable effect size was d = 0.74. Mean PcSource for the High search condition (M = .87, SD = .10) was significantly greater than mean PcSource for the Low search condition (M = .56, SD = .11), $t(43) = 9.56$, $p<.001$, d = 2.85, supported by a Bayes Factor, $BF_{10} = 3.19 \times 10^9$.

To investigate differences in clustering between the High search and Low search conditions a one-tailed t-test was run on ARC scores, collapsed across trials and Modality for the participants in the two search conditions. It was found that participants in the High search condition exhibited significantly greater ARC scores (M = .83, SD = .24) in their recall outputs, than participants in the Low search condition (M = .44, SD = .40), $t(43) = 4.06$, $p<.001$, d = 1.21, $BF_{10} = 235.39$. Therefore, there is strong evidence from both traditional and Bayesian statistics that participants who are better able to constrain search demonstrate a greater degree of clustering in their recall output, implicating an important role for context in constraining search.

The next main interest of the present study was whether participants could successfully monitor the output of search. These analyses were partitioned into target monitoring and source intrusion monitoring, as these 2 forms of monitoring are expected to differ, particularly at the start of the recall period. As previously stated, List membership is likely not a source associated with perceptual information. As this perceptual information is predominantly the basis upon which source judgments are made at the start of a recall period when retrieval is rapid, participants should not be able to monitor items retrieved early in the recall period. If this is true then target and source intrusion monitoring should be at chance at output position 1. Accuracy for both forms of monitoring should increase sharply with output position as retrieval

slows (both naturally and as a result of item-by-item reporting as previously explained), due to participants being able to access information upon which they can make source monitoring judgments.

Another possibility is that participants mostly neglect to monitor source early in the recall period, and make an assumption that the first item they retrieve must be a target. Indeed, Figure 2.5 demonstrates that search accuracy irrespective of List membership is extremely high at output position 1. If this source neglect account is true, then target monitoring accuracy should be at ceiling for output position 1 whereas source intrusion monitoring accuracy should be below chance. After this point participants will actively begin to engage in monitoring as source identifying information becomes available. Therefore, target monitoring should remain at or near ceiling for the entire recall period whereas source intrusion monitoring should progressively increase with output position. Neither form of monitoring is predicted to differ as a function of List membership. However, source monitoring may differ as a function of Modality if participants use serial order information to inform monitoring judgments.

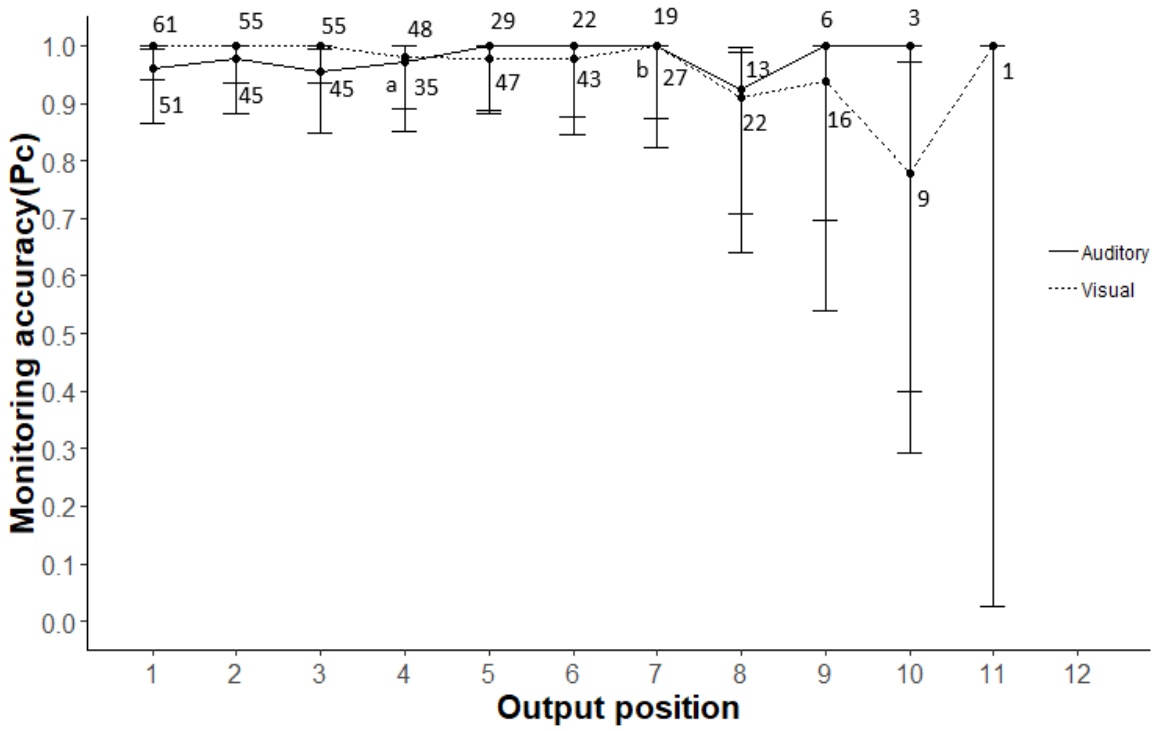
Single-sample *t*-tests were conducted to determine if participants could monitor targets both collapsed across Modalities and List membership, and broken down into individual Modalities. Minimum detectable effect size assuming .8 power for the fully collapsed analysis was $d = 0.38$, and $d = 0.55$ for the individual Modalities analyses. Collapsed across List membership and Modality, participants were able to monitor targets at above chance level, ($M = .97$, $SD = .06$), $t(44) = 53.29$, $p<.001$, $d = 7.94$. A Bayes factor demonstrates that there was extremely strong evidence for this, $BF_{10} = 2.01 \times 10^{38}$. Participants in the Auditory condition could monitor targets at above chance level ($M = .98$, $SD = .06$), $t(20) = 35.11$, $p<.001$, $d = 7.69$, $BF_{10} = 3.44$ x

10^{16}. The same was true of the Visual condition (M = .97, SD = .06), $t(23) = 39.27$, $p<.001$, $d = 8.02$ $BF_{10} = 3.69 \times 10^{19}$. The Bayes Factors for both Modalities indicate extremely strong evidence for above chance target monitoring. A one-tailed independent *t*-test revealed that target monitoring accuracy was not significantly higher for the Auditory modality than the Visual modality, $t(43) = 0.19$, $p = .42$, $d = 0.06$; however, the Bayesian evidence for the null was just short of being conclusive, $BF_{10} = 0.34$. Although the minimum detectable effect size for this analysis assuming .8 power was $d = 0.74$, it is likely that there is indeed no difference between the two Modalities given that an effect size of $d = 0.06$ is small.

Target monitoring dynamics were compared to see if there was a difference between the Modalities at any output position. BF_{10} for early (1-4), middle (5-8) and late (9-11) output positions were 0.35, 0.07 and 0.34 respectively. These Bayesian analyses demonstrate that there was not quite sufficient evidence to draw conclusions about target monitoring accuracy at the beginning and the end of the recall period although it is more likely that there is no difference. There was evidence for no difference between the Modalities at middle output positions. Target monitoring dynamics for the two Modalities are presented in Figure 2.6. It should be noted however that there were ceiling effects for target monitoring, which may mask potential differences between the Modalities.

Figure 2.6

Target Monitoring Accuracy by Output Position as a Function of Modality.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] Visual = 48 trials, Auditory 35 trials

[b] Visual = 27 trials, Auditory 19 trials

Next, single sample *t*-tests were conducted to observe if participants could monitor source intrusions at above chance level, both collapsed across Modality and List membership, and within each Modality. Minimum detectable effect sizes were the same as for Target monitoring. Collapsed across Modalities and List membership, participants could successfully reject source intrusions at above chance level (M = .83, SD = .28), $t(44) = 8.08$, $p<.001$, $d = 1.20$, $BF_{10} = 6.95 \times 10^{7}$. Separated into Modalities, participants were able to monitor source intrusions in the Auditory modality (M = .79, SD = .32), $t(20) = 4.13$, $p<.001$, $d = 0.90$ $BF_{10} = 127.74$, and the Visual modality (M = .87, SD = .23), $t(23) = 7.94$, $p<.001$, $d = 1.62$ $BF_{10} = 5.84 \times 10^{5}$. A one-tailed independent *t*-

test was then conducted to see if participants were better able to monitor source intrusions in the Auditory condition than the Visual condition. This was not found to be the case, $t(43) = -0.93$, $p = 0.82$, $d = 0.28$, $BF_{10} = 0.17$. Bayesian analysis shows very strong evidence for above chance source intrusion monitoring collapsed across Modalities, and partitioned into individual Modalities. There is also credible Bayesian evidence that source intrusion monitoring was not superior for the Auditory condition than the Visual condition.

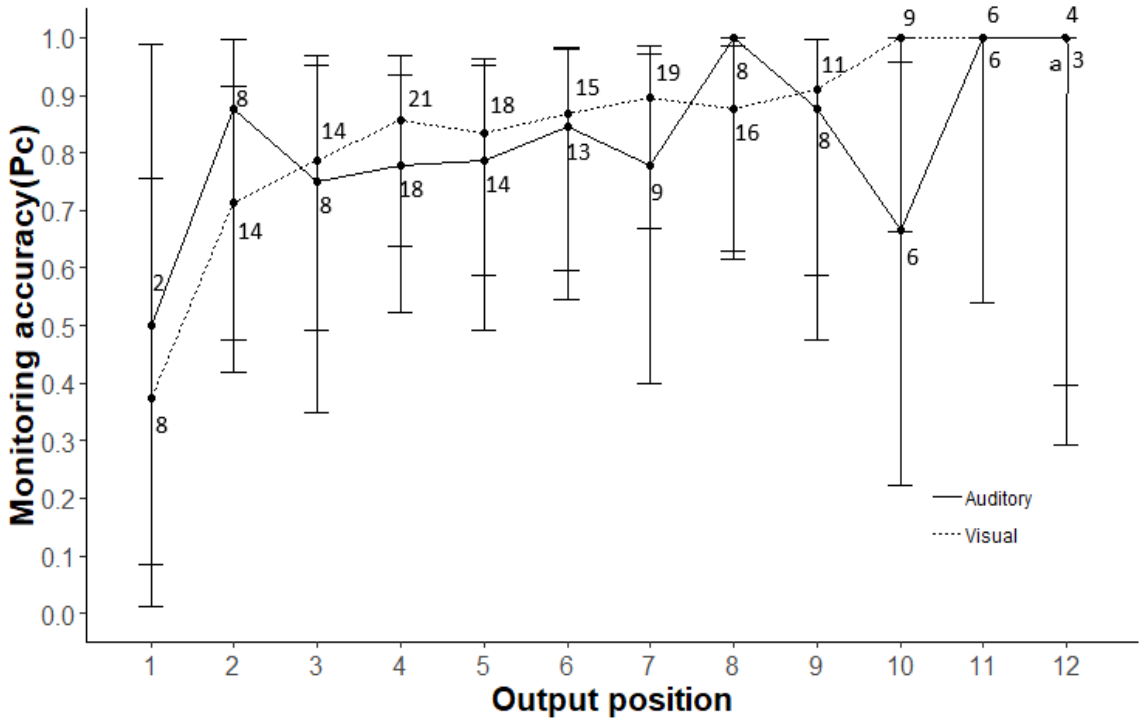
Source intrusion monitoring dynamics were compared across modalities to discern if there were differences at any stages in the recall period. BF_{10} for early (1-4), middle (5-8) and late (9-12) output positions were 0.25, 0.18 and 0.40 respectively. This indicates evidence that the modalities did not differ at early and middle output positions, but evidence for the null was inconclusive at late output positions. See Figure 2.7 for source intrusion monitoring dynamics. According to the source monitoring framework and CMR2, there are no predictions that the two lists should differ in source monitoring, so this was not followed up.

Finally, analyses were conducted to observe if there was a significant difference between target monitoring accuracy and source intrusion monitoring accuracy. A paired *t*-test (minimum detectable effect size assuming .8 power was $d = 0.38$) revealed that participants monitored targets with significantly greater accuracy than source intrusions, $t(44) = 3.27$, $p=.002$, $d = 0.71$. This was supported by a Bayes Factor, $BF_{10} = 15.30$. To investigate which points in the recall period this difference may originate from, monitoring dynamics were examined. BF_{10} for early (1-4), middle (5-8) and late (9-11) output positions were 8.50×10^9, 334.05 and 0.16 respectively. Output position 1 was examined in isolation to determine if source was being neglected, or whether participants were attempting monitor source but were unable to at this

output position. Source neglect would be characterised by evidence for a difference between target and source intrusion monitoring, whereas no ability to monitor would manifest as chance performance for both types of monitoring. BF_{10} for output position 1 was 7.30 x 10^4 indicating strong evidence for a difference and thus source neglect. By examining Figure 2.8 we can see that target monitoring is near ceiling for the majority of the recall period, whereas source intrusion monitoring accuracy is extremely poor at the beginning of the recall period, and improves with output position.

Figure 2.7

Source Intrusion Monitoring by Output Position as a Function of Modality.

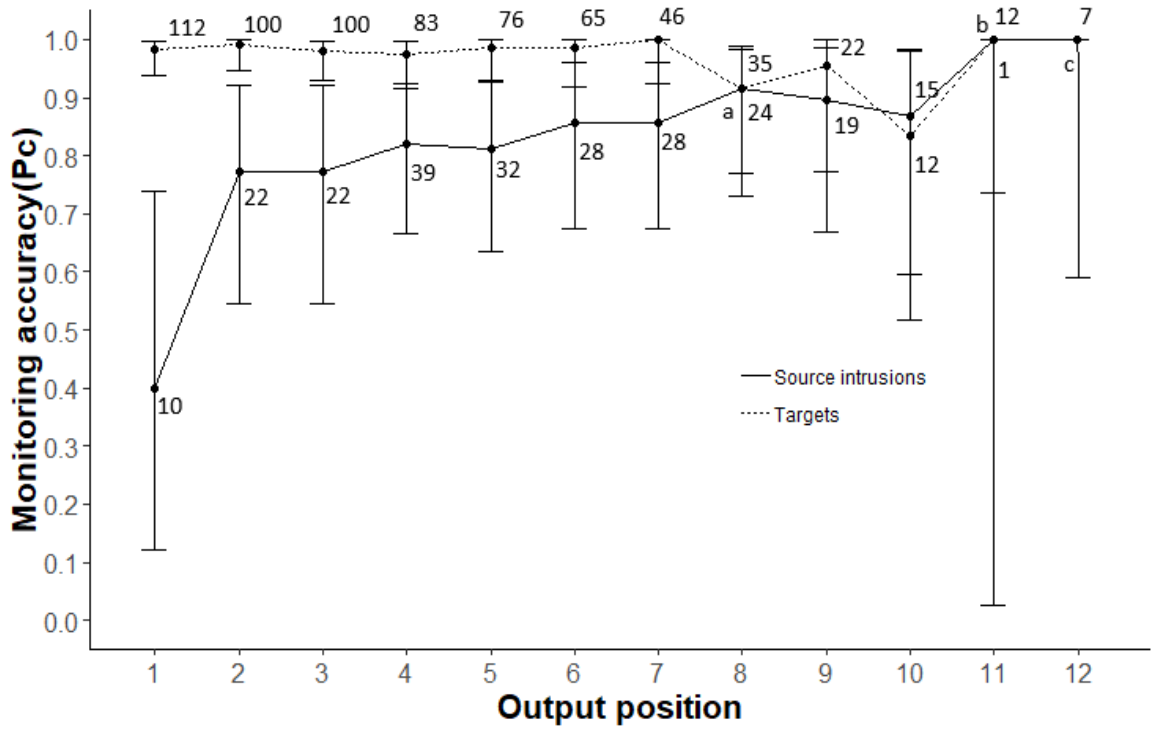


Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] Visual = 4 trials, Auditory = 3 trials

Figure 2.8

Target and Source Intrusion Monitoring Accuracy by Output Position.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] Targets = 35 trials, Intrusions = 24 trials

[b] Targets = 1 trial, Intrusions = 12 trials

[c] Only intrusions retrieved

2.4.3 - *Discussion*

This experiment first sought to determine whether search processes are affected by the unorthodox recall instructions of EFR. Comparison of recall and clustering across three different free-recall paradigms revealed that correct item availability and clustering were no different in EFR, when compared with two different free-recall paradigms with instructions to recall as many items as possible. Secondly, using EFR it appears that participants were able to constrain search to a target list of items. The general pattern of search dynamics appears to be a progressively reducing

search accuracy with output position. Furthermore, participants who were better at constraining search exhibited greater clustering in their recall outputs. This appears to implicate some role of context in constraining search, given that clustering is an indicator of contextually based search.

Participants were able to monitor both targets and source intrusions at above chance level; however, monitoring of these 2 response categories demonstrated very different patterns of accuracy throughout the recall period. Target monitoring remained at ceiling for almost the entire recall period. However source intrusion monitoring accuracy was extremely poor at the beginning of the recall period, rapidly increased at output position 2, and steadily improved thereafter. The dramatic difference between target and source intrusion monitoring at output position 1 is indicative of source neglect, manifesting in an extreme bias to identify the first generated item as a target without attempting to monitor source; hence, why target monitoring accuracy is so high yet source intrusion monitoring accuracy is less than 0.5. After this point, participants begin to actively engage in source monitoring, and the gradual increase in source intrusion monitoring accuracy with output position is consistent with retardation of retrieval, affording participants more time to assimilate information for more accurate source judgments. Finally, source monitoring did not differ as a function of Modality, implying that having better access to item serial order information does not improve the accuracy of source monitoring judgments.

2.5 - General discussion

The first main aim of the present chapter was to establish the new modified EFR paradigm as a viable and reliable method of simultaneously measuring constrained search and source monitoring. The main concern was whether the unorthodox recall instructions of EFR, namely to constrain recall to a subset of items, but also to write

down all incorrect source information that comes to mind whilst monitoring each retrieval, may affect search processes. Therefore, it was important that two measures of search processes; recall and clustering, were at least equivalent to paradigms where the instruction is to simply recall as many items as possible. Reassuringly, there was credible evidence that these two metrics in EFR were equivalent to both a standard-free recall experiment and an intermediate experiment where standard-free recall was paired with a monitoring instruction. Therefore, we can conclude that search processes in free recall are not hindered by the EFR recall instructions. Another issue associated with EFR is one of selective reporting, where participants deliberately fail to write down retrievals they suspect are incorrect; thus, artificially inflating constrained search scores. The following chapter will explore this issue further by investigating source contexts which should yield poorer constrained search accuracy than List membership.

The second aim was to utilise the new EFR paradigm to examine if participants can constrain search to a single list of items. A prediction derived from retrieved context models such as CMR2, (Lohnas et al. 2015) is that the cue for all retrievals irrespective of which list is the target list, is the current state of experimental context, which at the beginning of the recall period is the time-of-test context. The prediction is therefore that search accuracy for List 1 should be significantly worse than for List 2. This was not found to be the case as there was no difference in constrained search accuracy between the two lists. In addition there was little evidence that target or source intrusion availability differed as a function of List membership. Equivalent search performance across lists is important, as it suggests that participants do not simply use the current state of temporal context to retrieve all information, and that it

is possible to exclude recent information from search that is task irrelevant. Furthermore, analysis of search dynamics showed evidence that the two lists did not differ in constrained search accuracy throughout the recall period. One thing that cannot be established from the search dynamics alone, is whether the characteristic decrease in search accuracy over the course of the recall period is due to participants becoming less efficient at searching for targets later in the recall period, or increasing task difficulty as the recall period progresses, due to a gradual reduction in the base rate of novel targets. Chapter 5 will explore this in more detail.

The next aim was to investigate the role of context in constraining search. Clustering is long established phenomenon in the recall literature that is indicative of successful contextually-based search (Polyn et al. 2009a). It was hypothesised that if context plays a significant role in constraining search, then participants who are more efficient at constraining search should exhibit a greater degree of clustering in their recall output. There was strong evidence that this was indeed the case. Therefore, it is suggested that in order to retrieve a target list in the presence of another, participants may attempt to reinstate the encoded context of the target list at retrieval, by isolating items that have a strong contextual match with this reinstated context and excluding items that do not. However, for List membership studies it is not entirely clear what the nature of this reinstated context is. Although List membership can certainly be defined as a source, arguably there is no perceptual source information that distinguishes one list from another. Therefore source context may play no role in retrieving a target list. Instead the participant may attempt to reinstate the temporal context that was encoded at the time of presentation of the target list. The following chapter will explore in more depth the precise role of source context in constraining search by presenting Mixed-lists of items which do have distinguishing source

features.

The fourth aim of this chapter was to investigate whether participants can monitor the output of search at above chance level. Collapsed across output positions, participants were easily able to monitor targets, and were highly proficient at monitoring source intrusions. Monitoring dynamics yielded far more revealing information. The patterns of target and source intrusion monitoring can be explained very easily by the Source Monitoring Framework (Johnson et al. 1993). At output position 1 there was a vast disparity between target and source intrusion monitoring accuracy. Target monitoring accuracy was at ceiling whereas source intrusion monitoring accuracy was below 0.5. This suggests source neglect, where participants made no attempt to monitor source, and assumed that the first item recalled would be a target. Hence, why target monitoring accuracy was at ceiling and source intrusion monitoring accuracy was predominantly incorrect. Following this, participants actively engaged in monitoring, as evidenced by a dramatic increase in source intrusion monitoring accuracy by output position 2. The steady increase in source intrusion monitoring accuracy after output position 2 is also consistent with the Source Monitoring Framework. As previously stated, retrieval tends to slow exponentially as more items are output (Wixted & Rohrer, 1993); therefore, with each retrieval participants would have more time to assimilate source information to make more accurate source judgments.

The final aim was to investigate the potential role of incidentally encoded temporal context in retrieval, by presenting items to participants in either the Visual or Auditory modality. Serial-order information is better encoded in the Auditory modality (Neath & Crowder, 1990). Given that same source items in List membership experiments are also presented consecutively, serial order information could be useful

for constraining search and monitoring. It was found that Modality affected neither recall nor clustering in Experiments 2.1 and 2.2, indicating that serial order information was not influencing search. Upon reflection this was likely due to the nature of the task. Modality effects in recall tend to manifest as enhanced recency effects for the Auditory modality (Murdock & Walker, 1969); however, in delayed-recall tasks such as those presented in this thesis, recency effects are extinguished by the delay between item presentation and recall (Postman & Phillips, 1965). Therefore, participants are unlikely to derive benefit from the Auditory modality. As such, Modality effects were not investigated for constrained search. However, serial-order information may have been useful to participants in monitoring List membership. Ultimately there was little evidence that this was the case for distinguishing between targets and source intrusions in Experiment 2.3; however, evidence was too weak to draw conclusions about Modality effects on monitoring in Experiment 2.2.

The following chapter will examine in more detail the role of source context in constraining search and monitoring. In addition, factors which may influence the accuracy of these two processes will be investigated. This will also serve as a further validation of the modified EFR method as we can observe whether this paradigm is sensitive enough to detect predictable differences between contexts.

Chapter 3: Mixed-list contexts

3.1. Introduction

Chapter 2 demonstrates evidence for a constrained search mechanism, whereby participants can exclude incorrect source items from coming to mind. A modified version of EFR was also tested, and found to be a reliable and effective method of testing source constrained search and monitoring simultaneously, in the absence of potentially confounding prospective memory failures associated with the original button press procedure (Einstein & McDaniel, 1990). Despite finding evidence for constrained search, it is possible that this may not represent source based constrained search per se because of the contiguous nature of the sources in Chapter 2. Participants may be reinstating incidentally encoded temporal context rather than source context in order to retrieve a particular list. To solve this, the current chapter presents a series of experiments whereby sources are ordered randomly within lists. In this situation, sources are no longer related to serial position; therefore, temporal context is unhelpful in constraining search, and participants must actively attempt to constrain search by source. This is also an important test of the viability of modified EFR as a method of measuring constrained search and monitoring. If poorer constrained search is observed for Mixed-list contexts than for List membership as presented in Chapter 2, this is further evidence that participants are not selectively reporting items that they believe are correct. For modified EFR to be a viable method for measuring constrained search, this confound cannot be present.

Very few EFR studies have employed source manipulations which vary within a single list. One such study was Hollins, Lange, Berry and Dennis (2016), who explored

constrained search and monitoring in the context of cryptomnesia. Pairs of participants alternately generated exemplars of numerous semantic categories. After either one day or one week participants returned, and were given adapted EFR instructions to recall either their own or their partner's generated exemplars for all semantic categories, one at a time. Participants were required to write task compliant responses in one column, and anything else that came to mind in a separate column on a response sheet. This helped to avoid prospective memory failures associated with a keypress in standard EFR because an explicit monitoring decision was required for each item. It was found that source-recall errors were more common in the recall-partner task than the recall-own task. This was reflected in poorer constrained search and poorer monitoring in the recall-partner task. Specifically, own ideas were more likely to come to mind in both tasks, and own ideas were more likely to be reported incorrectly in the recall-partner than partner ideas in the recall-own task.

A further study by Hollins, Lange, Dennis and Longmore. (2016) examined social influences on unconscious plagiarism using EFR, namely the effects of the presence or absence of the partner at test. Similarly to Hollins, Lange, Berry and Dennis (2016), participants generated exemplars to semantic categories alternating with their partner. They were told not to duplicate either their own previous ideas or any of their partner's ideas. A week later participants returned either with their partner, or alone, for EFR for either their own or their partner's ideas.

Firstly, the presence or absence of a partner at recall had no effect on generation of their partner's ideas. The only significant effect of partner presence was in the recall-partner task, where participants were more likely to generate task-irrelevant own ideas if the partner was present. There was however one clear effect for monitoring. Irrespective of the recall task, participants were less likely to report

wrong source items as task compliant when the partner was present. This was attributed to greater attention being paid to source when the partner was present, reducing the likelihood of source neglect. Partner presence did not seem to shift response bias as there was no overall increase in the number of correct source items or source intrusions reported as task-compliant.

In order to ascertain which within-list source manipulations should be utilised for the current chapter, it is important to examine the source monitoring literature to gain an insight into which source dimensions participants can reliably differentiate. Participants can successfully retrieve a variety of intrinsic stimulus features in source recognition tasks, such as the colour of a word and its location on a computer screen (Mulligan, 2004), font size (Starns & Hicks, 2005) and modality (Hintzman et al., 1972). Source retrieval of external properties of a stimulus such as background context is equivocal. Doerksen and Shimamura (2001) found chance performance for source recognition of background border colour. Although source monitoring was significantly better for emotional than neutral words. This was reflected in far greater recall for emotional words than neutral words, suggesting that source recognition was related to the memorability of the item. However, using a very similar border colour manipulation but with pictorial stimuli, Ecker et al. (2007) found significantly above chance source recognition for background border colour. In fact, there was no significant difference between this source manipulation and recognition for object colour. This may be reflected in their use of pictures as stimuli. It has long been established that pictures are remembered better than words, an effect known as picture superiority (Mintzer & Snodgrass, 1999; Paivio & Csapo, 1973). If recognition of external sources is related to item memorability, it is unsurprising that Ecker et al.

(2007) observed above chance source recognition where Doerksen and Shimamura (2001) failed, as the former study used more memorable stimuli.

The present chapter had a number of aims. The first was to investigate potential selective reporting confounds in EFR. For List membership, participants are likely to reinstate temporal context in order to constrain search given that the only distinguishing feature between the two lists is time. CMR (Polyn et al. 2009a) states that this same temporal context would then be used to guide retrieval. In this case items from the same list are likely to be activated much more strongly during the recall period than items from a different list. If participants successfully reinstate the target context, then long runs of target retrievals may be observed simply by virtue of their consecutive presentation. Participants may then stop retrieving when they cannot generate any more items from the target list.

For Mixed-list experiments, the two candidate sources are presented in a random order, and therefore temporal context is unhelpful in retrieving targets. In this instance the sources do contain distinguishing perceptual information, so the retrieval cue comprises source context and temporal context, but only source context is useful in retrieving targets. In some cases temporal context may increase the likelihood of wrong-source items being generated if they were presented proximally to a target. Therefore, CMR predicts that constrained search should be poorer for Mixed-list source manipulations than for List membership. If EFR does not suffer from selective reporting confounds, the paradigm should be sensitive enough to detect such predictable differences in constrained search between different forms of context.

The second aim was to examine differences in monitoring between List membership and Mixed-list contexts. The prediction made by the Source Monitoring Framework (Johnson et al. 1993) is that, given that there is rapidly accessible

perceptual source information associated with Mixed-list contexts, which is not present for List membership, the tendency to neglect source should be less for Mixed-lists than for List membership. This should manifest as superior source intrusion monitoring for Mixed-lists than List membership early in the recall period, particularly at output position 1.

The third aim was to examine factors which may influence the effectiveness of constrained search and source monitoring using the modified EFR paradigm. The first experiment examines the effects of Source Similarity. The Source Monitoring Framework states that two sources which are highly similar, are more difficult to differentiate than sources which are less similar. Therefore, source monitoring judgments are less accurate in the High-similarity scenario. It was predicted that the same principle would apply to constrained search. When the target source and the wrong source are highly similar, constrained search should be less successful than if the target and wrong source are less similar. In fact, this exact prediction is also made by CMR.

The other factor which may affect constrained search that this chapter examines is Context Dependency. The Source Monitoring Framework makes the prediction that source monitoring judgments become more accurate as more source information is made available to the participant. This framework and CMR also make the prediction that constrained search should also be more successful when there is more source information available at the start of the recall period. To explore this, an experiment was run where additional source information afforded to the participant was either helpful or unhelpful in retrieving targets. Dependency refers to the probability that a source from one studied context predicts the source of another context. An example of Full Dependency would be if all items

printed in red were also printed in large font, and all items printed in blue were printed in small font. Full independence would be if both colour and size were assigned to the item randomly; therefore, the 2 contexts do not predict each other. In the Full Dependency case, participants can for example use source information about an item's size in order to constrain search by colour, and vice versa whereas this is not possible for the Full Independence case. Therefore, constrained search should be more successful in the Full Dependency case. The same would apply to source monitoring, where inferences can be made about one source from a different context in the case of Full Dependency, increasing the accuracy of monitoring judgments.

3.2 - Experiment 3.1

One of the key findings from the previous chapter was that participants who were better able to constrain search exhibited greater clustering in item generation, implicating a role for context in constraining search. Therefore clustering can be used as a marker for which source contexts participants should be able to search by, and which factors should influence constrained search.

A body of evidence suggests that clustering is driven by patterns of similarities and differences among items. Frost (1971) presented participants with thirty-two line drawings in one of four orientations. They were then asked to recall the names of the pictures. Other participants were presented with the names of these same line drawings, with no manipulation of orientation. Clustering was scored according to organisation by orientation category. For the pictorial stimuli, significant clustering by orientation was observed. This was not the case for the verbally presented stimuli. This evidence suggests that participants searched items by visual similarities among the line drawings. The fact that the same items did not cluster in a similar fashion in their

verbal form, indicates that the orientation clustering cannot be attributed to semantic relatedness within the orientation categories.

One of the most commonly used source manipulations when investigating Source Similarity is study modality. Hintzman et al. (1972) presented three experiments whereby they manipulated source both across modality and within modality. For each study, participants were presented with eight lists of eighteen words followed by an immediate-free-recall test. For Experiment 1 the source manipulation was visual vs auditory presentation. Experiment 2 manipulated source by presenting items in two different fonts, and the source manipulation for Experiment 3 was voice gender. Effectively there was one across-modality manipulation and two within-modality manipulations, visual and auditory respectively. The implication here was that the two sources employed in Experiment 1 were less similar than the two sources presented in the other experiments. Significantly above chance clustering was observed in all three experiments. Interestingly, there was a greater degree of clustering in Experiment 1 than either Experiment 2 or 3. This suggests a greater degree of clustering when sources are more dissimilar. Furthermore, source recognition was superior for the across modality manipulation than for either within modality manipulations. Rates of correct source identification were 74% for across modality, 58% for within visual and 59% for within auditory. This study clearly demonstrates the influence of Similarity on both search and old/new source recognition after the recall period.

Further evidence for organisation of memory by similarity was presented by Nilsson (1974). Participants were presented with Mixed-lists of items presented in two sources. In one condition, the two sources were visual and auditory presentation, the second condition was male and female voices and the third was uppercase and

lowercase letters. Again, this constituted one across-modality and two within-modality conditions. Immediate-free recall followed presentation of each list. Participants were asked to recall as many items as they could, but also to identify the source of each item as it was recalled. Similar to Hintzman et al. (1972), it was found that clustering by Modality was superior to clustering by either of the within-modality conditions. However, source monitoring was equal for all three types of list presentation. The discrepant findings in source monitoring for across versus within modality manipulations may originate from the way source monitoring was tested. Monitoring in the Nilsson (1974) study was online, i.e., during the recall period, whereas for Hintzman et al. (1972), source monitoring was tested offline by source recognition after the recall period. The Source Monitoring Framework (Johnson et al. 1993) acknowledges that source monitoring can occur due to source neglect, whereby the participant does not engage in monitoring. Indeed, this was evident in Experiment 2.3. This is far more likely to occur during online source monitoring where participants have limited time to make source judgments, particularly at the beginning of the recall period. For offline source monitoring as per Hintzman et al. (1972) participants have time to accumulate sufficient source information for each judgment, leading to higher overall source monitoring accuracy.

The present experiment had three overarching aims. The first was to assess the suitability of EFR for measuring constrained search given its potential for selective reporting. The best way to do this is to compare different source contexts that should result in significantly different constrained search scores. In this instance, I will compare constrained search from List membership (Experiment 2.3) with Mixed-lists of different modality manipulations. Retrieval models such as CMR (Polyn et al. 2009a) state that temporal context, for example inter-item associations is necessarily encoded

with an item in addition to its source. At retrieval for List membership, this incidentally encoded temporal context will likely only increase activation of target items as targets are temporally proximal. However, for Mixed-lists temporal context will likely increase the activation of the incorrect source, as the two sources are randomly ordered. This ultimately means that more incorrect items will be included in the search set for Mixed-lists than for pure lists. In addition, incorrect items are more likely to be brought to mind in Mixed-list experiments. Therefore, if EFR is suitable for measuring constrained search, then superior constrained search for List membership than Mixed-list contexts should be observed.

The second aim is to ascertain whether Source Similarity has a significant effect on constrained search, as this is yet to be investigated in the literature. Participants will study items in High-similarity trials (Male vs Female voices) and Low-similarity trials (Auditory vs Visual presentation). CMR implies that in the High-similarity trials, the target-source retrieval cue will be less distinctive than in the Low-similarity trials. Therefore, incorrect items will receive a greater degree of activation from the cue than incorrect source items in the Low-similarity trials. One would expect more incorrect items in the search set for the High-similarity trials, hence poorer constrained search accuracy overall. If no differences are observed between the Similarity conditions collapsed across all output positions, differences may still be seen in the search dynamics. The purest measure of retrieval cue distinctiveness is search accuracy at output position 1, as the first retrieval is unaffected by inter-item temporal associations (temporal context) that drive search. If the retrieval cue in the High-similarity lists is less distinctive than that in the Low-similarity trials, then search accuracy in the High-similarity trials should be lower at output position 1 than the Low-similarity trials.

The third aim is to assess the impact of Source Similarity on source monitoring. A very simple prediction can be made from the Source Monitoring Framework (Johnson et al. 1993). This framework states that the success of source judgments is based on a number of factors. These include, but are not limited to, the amount and type of source information within the memory trace, the similarity of two candidate sources and the criterion the participant uses to make their judgment. The framework simply posits that source monitoring judgments are less accurate in situations where sources are more similar than less similar. In the context of the present study, this means that source monitoring should be superior for the Low-similarity trials than the High-similarity trials.

3.2.1 - *Methods*

3.2.1.1 - *Participants*

Sixty-four Psychology undergraduates took part in this study (11 Male, 53 Female, Mean age = 20.52, *SD* = 3.51), in return for compulsory course credit to pass a module.

3.2.1.2 - *Design*

There was one within-subjects manipulation for this study, Source Similarity. Eighty words were randomly allocated to one of four twenty-item experimental trials. In the Low similarity condition, participants were presented with two trials of twenty words. Half of the words in each list were presented auditorily, through headphones in a male voice, and the other half were presented visually on the computer screen. A thirty second numerical distractor task appeared after the tenth and twentieth word of each trial for consistency with Experiments 2.1-2.3 described in the previous chapter. There were five visual, and five auditory words before and after the first numerical distractor presented in a random order.

In the High-similarity condition, all stimuli were presented through the headphones. Half of these were spoken in a male voice, and the other half in a female voice. As with the Low similarity condition, the trial was divided into two lists by the numerical distractor task, and an equal number of male and female presented words appeared in each list, with a random order. After each trial had finished, the distractor task appeared a second time. Participants completed two trials per Similarity condition. Memory was tested four times. Each test occurred thirty seconds after presentation of the second list of each trial.

The order of conditions was counterbalanced by participant number so that half the participants completed the High-similarity trials followed by the Low-similarity trials, and the other half received the opposite condition order. The precise order of recall within these conditions i.e. recall auditory/recall visual was also counterbalanced by participant number. Allocation of participant numbers to condition order and recall order was determined prior to the experiment by means of random sampling without replacement.

3.2.1.3 - *Materials*

Stimuli for the experimental trials were eighty concrete nouns drawn from the updated and expanded Battig and Montague (1969) norms (Van Overschelde et al., 2004). These comprise exemplars of seventy semantic categories. Each exemplar is presented with information regarding its geographical and generational stability, and likelihood of generation. Using this pool allowed greater control over semantic relatedness among stimuli. This was achieved by selecting eighty items from across all seventy categories. This would hopefully reduce (although not eliminate) the potential for consecutive generation of items purely due to semantic relatedness.

Visual stimuli in the Low-similarity condition were presented on a computer screen in black Courier New font; size 32; against a white background. For all auditory stimuli in both Similarity conditions the computer screen was blank throughout stimulus presentation. Auditory recordings of all eighty words were made in a real human male voice, and in a real human female voice. Both voice actors spoke each word into a recording device. Audio recordings were made in such a way that each stimulus was represented by a single audio file (.WAV). The voice used for auditory stimuli in the Low-similarity condition was the same as the male voice in the High-similarity condition. Stimuli were allocated to trials such that no individual word appeared in more than one trial for any given participant. For instance the word 'horse' could not appear as an auditory stimulus in the Low-similarity condition, and again as female spoken word in the High-similarity condition for the same participant.

To account for individual differences in hearing ability, volume was adjusted manually to suit the participant prior to the experiment, by presenting a series of beeps of different volumes through headphones. Participants were asked to indicate which was the loudest volume that they were comfortable with. Before running any experimental participants, two individuals who were not involved in the study piloted the auditory stimuli to check that they were comprehensible, and represented the intended words. All eighty stimuli in both voices were presented to these pilot participants through headphones. For each stimulus they were asked to write down the word they heard on a sheet of paper. All words reported by the pilot participants matched the respective auditory stimuli.

Stimuli for the practice trials were ten Snodgrass and Vanderwart (1980) images in their original pictorial form. These were presented in the centre of the

screen, with a height and width of 75% screen size. All participants received the same practice stimuli.

3.2.1.4 - *Procedure*

Before the first trial of the Low-similarity condition, participants were told that for each trial, they should remember as many words as they could from both lists (forgetting words from previous trials), and to remember how each word was presented (through the headphones or on the computer screen). They were then presented with five words through the headphones (auditory), and five words in the centre of the computer screen (visual), in a random order one at a time. Each word had a presentation duration of four seconds, with a two second inter-stimulus interval (ISI). It should be noted that for all auditory stimuli irrespective of Similarity condition, each spoken word was of different duration; therefore, stimuli were presented over a four second period, with silence filling time when words were not being spoken. Following the tenth word, participants completed the numerical distractor task as described in the previous chapter. Then, the final ten auditory and visual words (five auditory, five visual) were presented randomly. The numerical distractor task then appeared again for thirty seconds (black text; Courier New font; size 32; white background). The High-similarity condition was almost identical. Study instructions were that participants should remember as many words as they could from both lists (forgetting all previous trials), and to remember how each word was presented (in a male voice or a female voice). The only difference in stimulus presentation was that all stimuli were presented through the headphones; half of the words in a male voice and the other half in a female voice. For both conditions, following the second numerical distractor task a screen appeared stating 'Choose A or B'. Participants were told that these letters corresponded to one of the two sources used in that list (auditory or

visual for Low-similarity and male or female for High-similarity) selected at random. This was employed so that participants could not expect to recall a particular source for a given list, and therefore, only attend to items presented in that source at study. Once participants had pressed either the A or B key on the keyboard, a screen appeared instructing them to recall items presented in one of the two sources used in that list.

At this point, a tablet device was given to the participant. The appearance and functionality of the tablet screen was identical to Experiment 2.3. Participants were required to recall only the items indicated by the computer screen, and to write them using a stylus in the 'target' box. However, if any items presented in the opposing (incorrect) source happened to come to mind, then they were written in the 'other' box. After participants finished writing a word, they pressed a 'next' button in the top right corner of the box to clear it for the next item. Previously recalled items were invisible to the participant. When the participant could not remember any more words from the source indicated on the computer screen, they pressed the finish button. After a four second blank screen, the 'target' and 'other' boxes reappeared. Participants were then required to perform the same recall task as the first recall attempt; however, the target and incorrect sources were switched so that participants needed to recall only the items presented in the previously incorrect source. Although ultimately the second recall attempt was not analysed due to potential carry-over effects, it was retained (though not analysed) in order that the procedure matched that of Experiment 2.3 as closely as possible, given that search and monitoring performance across those experiments would be contrasted. After this, they pressed the spacebar on the keyboard, and a screen appeared saying 'Are you ready for the

next list'? Pressing space again started the following trial. After the second trial, participants were informed that the sources used in the third and fourth trials would be different, and what these would be (see above for description of instructions for both conditions). Following the fourth trial, a screen stating 'The experiment is now over. Thank you for your participation' appeared.

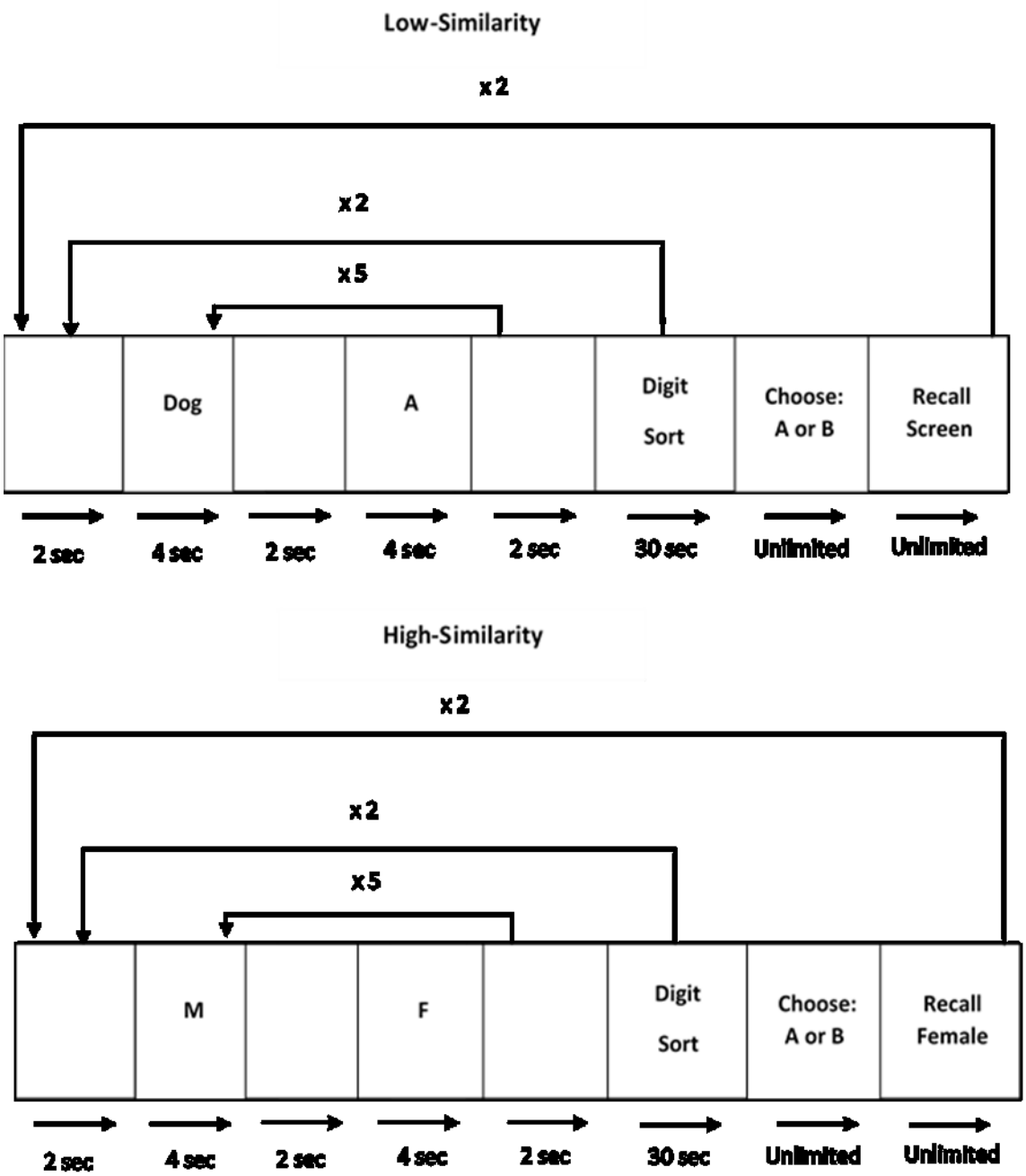
Prior to the first trial, participants underwent a practice trial to familiarise the them with the EFR procedure. This used pictorial stimuli instead of words to minimise interference with the experimental trials. Participants studied ten pictures, five either side of the first numerical distractor task. EFR instructions were to recall only the items from List 1 (before the first numerical distractor) or List 2 (after the first numerical distractor), as indicated by the computer screen, in the same fashion as described for the experimental trials. See Figure 3.1 for a schematic representation of the experimental paradigm.

3.2.1.5 - *Scoring*

Scoring was identical to Experiment 2.3.

Figure 3.1

Schematic Representation of Paradigm for Experiment 3.1.



Note. Digit sort = Digit sorting distractor task used throughout this thesis. A = Auditory stimulus in Low-similarity condition. M = Male voice in High-similarity condition. F = Female voice in High-similarity condition. This representation demonstrates Low-similarity condition first. In reality, half of the participants completed this condition second, as condition order was counterbalanced by participant number. In addition, participants did not all recall the same source in each condition.

3.2.2 - *Results*

3.2.2.1 - *Comparison with List membership*

As comparisons were being made between two experiments whose participants were drawn from different populations (undergraduates for Experiment 3.1 and general public for Experiment 2.3), it was necessary to investigate whether age differences between these populations could at least partially account for any differences in constrained search and monitoring between the two contexts. An independent *t*-test revealed that participants in the List membership experiment were significantly older (M = 23.17, SD = 4.52) than participants in the present study (M = 20.52, SD = 3.51), $t(110) = 3.49$, $p<.001$, $d = 0.67$. This was supported by a Bayes Factor, $BF_{10} = 40.36$. Although there was good evidence for a difference in ages between the populations, and a moderate effect size, it is unlikely that age in reality could complicate interpretation of the data, given that the difference in mean ages between the two populations was only 2.65 years.

Before Source Similarity was examined, it was first important to establish whether selective reporting was a significant confound in the modified EFR procedure. If this is an issue, any differences in constrained search due to the Similarity manipulation may be masked by participants deliberately not reporting wrong-source items they have generated. Therefore, constrained search (and monitoring) from this experiment will be compared with Experiment 2.3. As previously explained, the prediction would be that if participants are responding appropriately to the task instructions, then constrained search should be significantly poorer in the present experiment than for List membership (Experiment 2.3). Analysis of Source Similarity will follow.

Collapsed across Similarity conditions, in the present experiment, participants

generated a mean of 4.53 targets (*SD* = 1.68), compared with 5.16 targets for List membership (*SD* = 1.99). Independent *t*-tests were conducted to investigate whether target availability, source intrusion availability and overall search accuracy (PcSource) significantly differed between the present experiment and List membership context (Experiment 2.3). For assumed power of .8, minimum detectable effect size was *d* = 0.53. There was no significant difference in target availability between the two contexts, $t(107) = 1.77$, $p = .08$, $d = 0.34$; however, a Bayesian *t*-test revealed that there was very little evidence at all for this, $BF_{10} = 0.83$. For Mixed-lists, participants generated a mean of 2.45 source intrusions (*SD* = 1.48), compared with 2.06 source intrusions for List membership (*SD* = 1.63). There was no significant difference in source intrusion generation between the two contexts, $t(107) = 1.30$, $p = .20$, $d = .25$, however the Bayesian *t*-test was inconclusive, $BF_{10} = 0.44$. Therefore, unfortunately, due to low power and inconclusive Bayes Factors we cannot draw firm conclusions about differences between search in Mixed-lists and List membership from numbers of targets and source intrusions alone.

PcSource scores were investigated to observe if participants could constrain search at above chance level (0.5) for Mixed-lists. A single sample *t*-test (minimum detectable effect size, $d = 0.31$) revealed that this was the case, $t(63) = 7.70$, $p<.001$, $d = 0.97$, supported by a Bayes Factor, $BF_{10} = 8.41 \times 10^7$. A one-tailed *t*-test (minimum detectable effect size, $d = 0.48$) was conducted to observe if there was a difference in constrained search between the contexts; the hypothesis being that PcSource should be significantly higher for List membership than Mixed-lists. The *t*-test revealed that this was the case, $t(107) = 1.87$, $p = .03$ $d = 0.36$, however the Bayesian evidence was inconclusive, $BF_{10} = 1.84$. Given that the effect size is also lower than the minimum detectable effect size, this implies that the effect can still be seen, although less than

four times out of five when the alternative hypothesis is true. See Table 3.1 for descriptives.

Table 3.1

Targets and Source Intrusions Generated, Overall Search Accuracy and Proportions of Targets and Source Intrusions Monitored Correctly for Mixed-lists and List Membership.

Measure	Mixed-list		List membership	
	M	*SD*	*M*	*SD*
Targets	4.53	1.68	5.16	1.99
SI	2.45	1.48	2.06	1.63
PcSource	**0.65***	0.17	**0.72***	0.19
T mon (Pc)	**0.99**	0.04	**0.97**	0.06
SI mon (Pc)	**0.84**	0.19	**0.83**	0.28

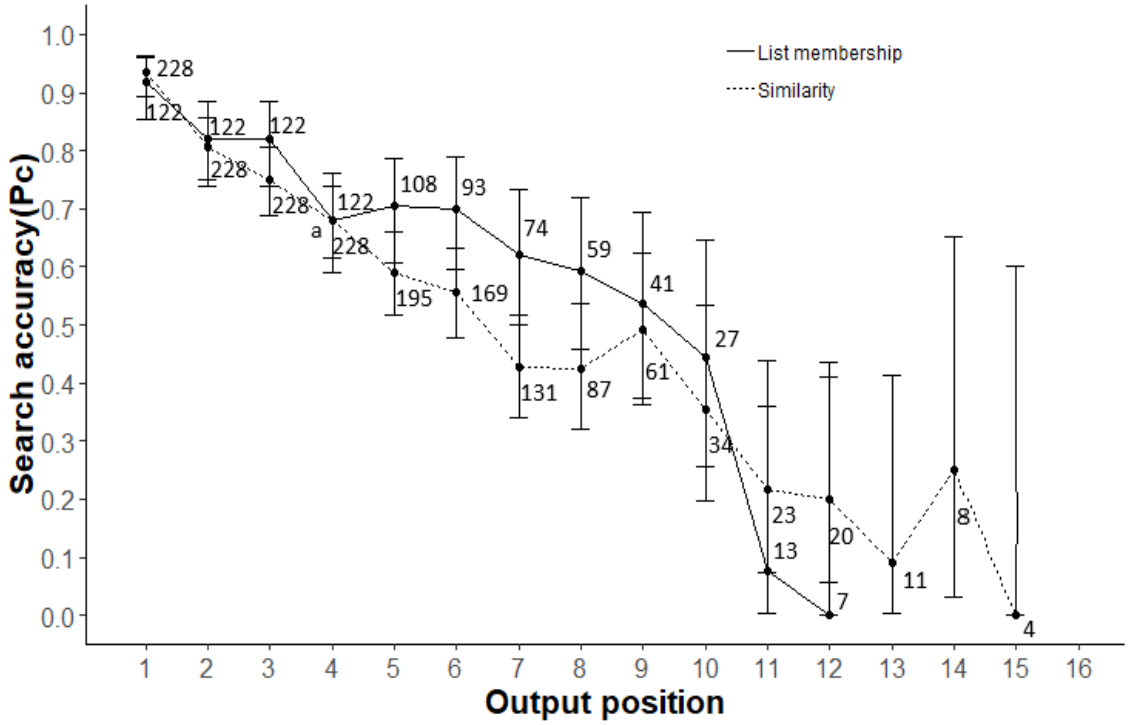
Note. Mixed-list scores were collapsed across Similarity condition. Bold text indicates significantly above chance performance. SI = Source Intrusion, T mon = Target monitoring, SI mon = Source intrusion monitoring. M = mean, SD = Standard Deviation.

*p < .05

Bayesian contingency tables were conducted on each bin to investigate differences between the two contexts across the recall period. BF_{10} for early (1-4), middle (5-8) and late (9-12) output positions were 0.07, 920.47 and 0.18 respectively. Output position 1 was examined in isolation to determine if there were differences in distinctiveness between temporal and Mixed-list context cues. BF_{10} for output position 1 was 0.09, indicating that there was no difference in cue distinctiveness between the two contexts. See Figure 3.2 for search dynamics. The trend for List membership appears to be a gradual decline in search accuracy followed by a precipitous drop near the end of the recall period, around output position 10. The Mixed-list trend is characterised by a steeper decline in search accuracy after output position 4. The difference between the two contexts narrows at output position 9. After this, accuracy for Mixed-lists is superior than for List membership.

Figure 3.2

Search Accuracy by Output Position for List Membership and Mixed-List (Similarity) Contexts



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] List membership = 122 trials, Similarity = 228 trials.

Given that the auditory source in the Low-similarity condition and the male source in the High-similarity condition were identical (the same male voice), it was important to examine whether there was interference across conditions. For example, a participant may retrieve an auditory source item presented in trial 2 (Low-similarity condition) during a trial 3 (High-similarity) recall period, believing that the trial 2 auditory source item was presented as a male source item in trial 3. To investigate this, each participant's trial 3 and trial 4 recall outputs were examined for instances of items presented in the same male voice in trials 1 and 2 (either the auditory source or male source depending on the condition order). The total number of these intrusions across trials 3 and 4 were then divided by the total number of generated items in trials 3 and

4, which yields the proportion of these interference intrusions in the recall outputs. This is expressed in Equation 3.1.

$$P_{int} = \frac{Int_3 + Int_4}{n_3 + n_4} \quad (3.1)$$

Int_3 and Int_4 are male voice interfering items in the trial 3 and 4 recall outputs respectively, and n_3 and n_4 are the total number of generated items in trial 3 and trial 4 respectively. A P_{int} score that is significantly greater than 0 would indicate that the second condition in this paradigm is partly confounded by interference from one of the sources in condition 1. Mean P_{int} for the present experiment was .003 (*SD* = .01) A single sample *t*-test (minimum detectable effect size for .8 power was *d* = 0.31) revealed that this was not significantly greater than 0, $t(63) = 1.52$, $p = .06$, $d = 0.19$; however, the Bayes Factor was inconclusive, $BF_{10} = 0.76$. This analysis is potentially underpowered although the effect size is small, and in reality only two of the sixty-four participants in the sample generated such interference intrusions. Therefore, it would seem that this type of interference should not be seen as concerning.

Regarding source monitoring, for List membership there was evidence for source neglect, where participants were predominantly monitoring the first generated item as a target irrespective of whether the item was a target or a source intrusion, indicating that they had not monitored the item at all. The Source Monitoring Framework (Johnson et al. 1993) predicts that source neglect should be far less prominent in Mixed-lists, as there is perceptual source information that distinguishes the two sources with Mixed-lists, whereas this is not the case for List membership. This perceptual information may be available immediately upon retrieval of the item, so it is more likely that participants will be able to make a source judgment based on some form of source identifying information. This is not to say that source neglect will be

absent for Mixed-lists, as some participants may still make an assumption that the first generated item is a target; however, this should be less pronounced. Ultimately target monitoring should be either at or near ceiling level for Mixed-lists, similar to List membership. On the other hand, source intrusion monitoring should be superior for Mixed-lists, given the additional source information available particularly at early output positions.

For assumed power of .8, the minimum detectable effect size for one-sample *t*-tests was $d = 0.31$ and $d = 0.48$ for an independent *t*-test. The proportion of targets correctly accepted for Mixed-lists was .99 (*SD* = .04), which was significantly above chance, $t(63) = 101.38$, $p<.001$, $d = 12.67$, $BF_{10} = 6.86 \times 10^{67}$, and .97 (*SD* = 0.06) for List membership, again significantly above chance $t(44) = 53.29$, $p<.001$, $d = 7.94$, $BF_{10} = 2.01 \times 10^{38}$. The Bayes Factors demonstrate very strong evidence for above chance performance in both contexts. There was no significant difference in target monitoring between the two contexts, $t(107) = 1.42$, $p = .16$, $d = 0.28$, although this may be due to lower power in the analysis. In addition the Bayesian evidence was inconclusive, $BF_{10} = 0.51$.

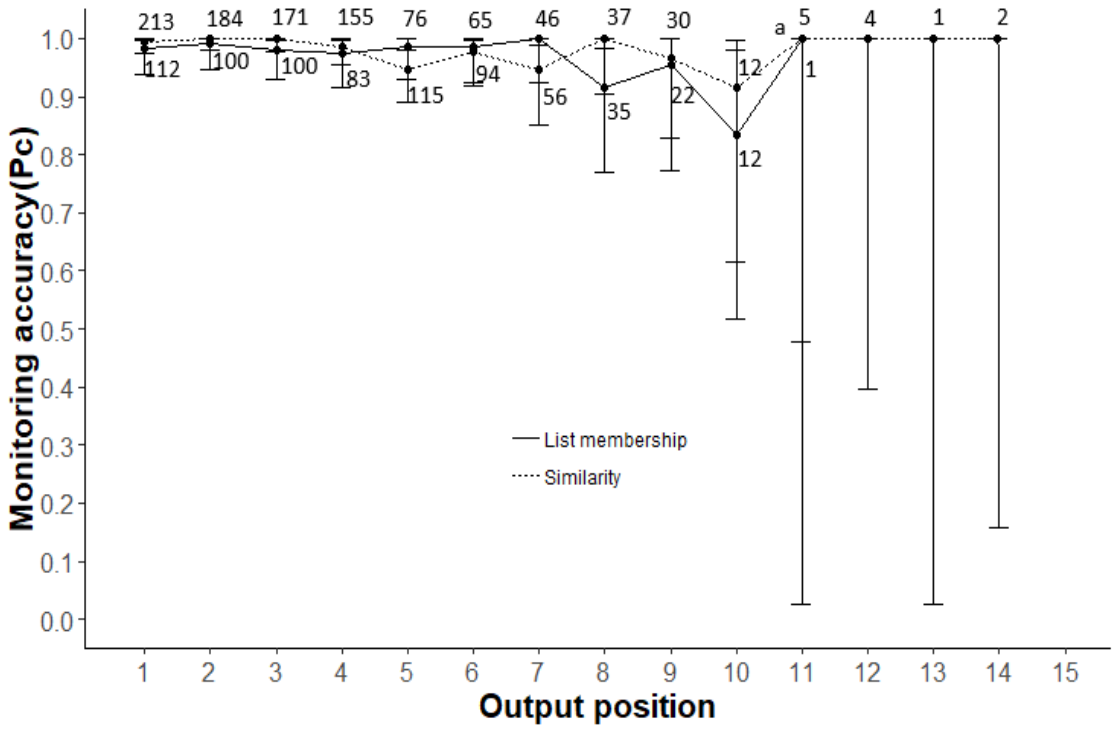
For target monitoring dynamics BF_{10} for early (1-4), middle (4-8) and late (9-11) output positions were 0.18, 0.06 and 0.19 respectively. The trend appears to be very high target monitoring accuracy throughout the recall period, with small reductions for List membership at output positions 8 and 10 (see Figure 3.3); however, it is unknown why these output position alone would yield slightly poorer target monitoring accuracy.

Minimum detectable effect sizes for source intrusion monitoring were the same as for target monitoring. The proportion of source intrusions correctly rejected for Mixed-lists was .84 (*SD* = .27), which was significantly above chance, $t(63) = 10.03$,

p<.001, d = 1.25, BF_{10} = 1.28 x 10^{12}, and .83 (*SD* = .27) for List membership, again significantly above chance, $t(44) = 8.08$, p<.001, d = 1.20, BF_{10} = 6.95 x 10^{7}. These Bayes Factors indicate very strong evidence for above chance performance. A one-tailed *t*-test revealed that source intrusion monitoring was not significantly more accurate for Mixed-lists than for List membership, $t(107) = 0.08$, p = .47, d = .01, supported by a Bayes Factor BF_{10} = 0.22. This does not seem to be an issue with power as the effect size is very small; therefore, it is most likely that source intrusion monitoring did not differ between the contexts.

Figure 3.3

Target Monitoring Accuracy by Output Position for List Membership and Mixed-List (Similarity) Contexts.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] List membership = 1 trial, Similarity = 5 trials.

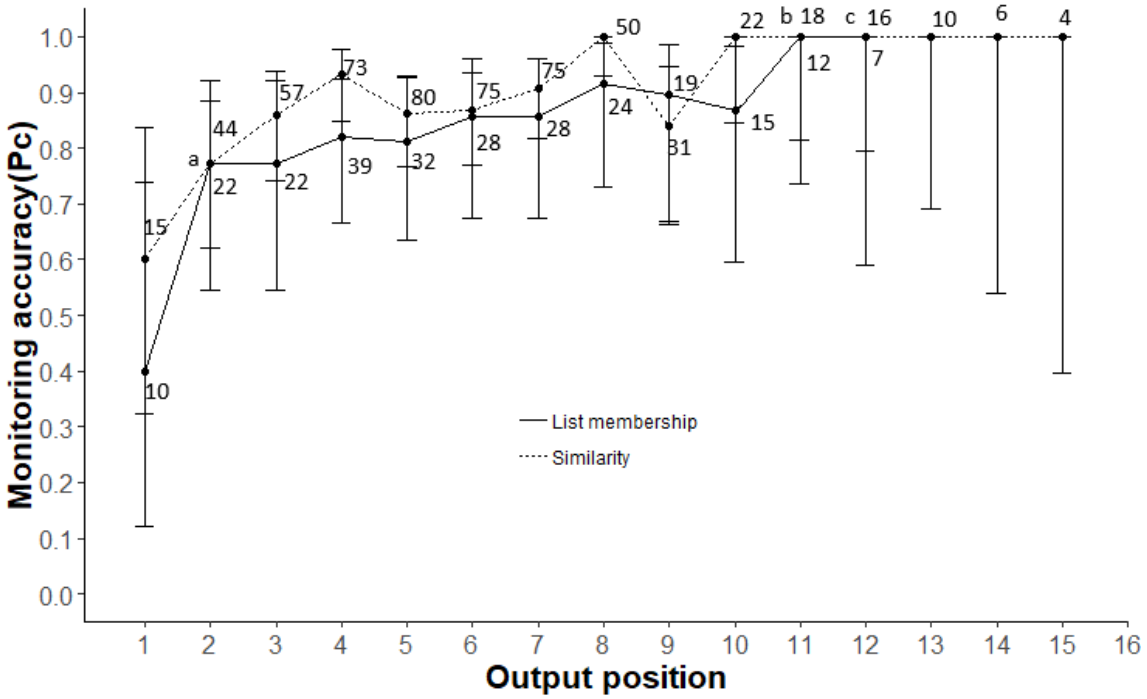
The same method that was used to analyse target monitoring dynamics was

employed to investigate differences in source monitoring dynamics between the two contexts. BF_{10} for early (1-4), middle (5-8) and late (9-12) output positions were 0.74, 0.19 and 0.12 respectively. This implies weak evidence for no difference between the contexts early in the recall period, and credible evidence for no difference after this. There is a notable difference between the contexts at output position 1, which would indicate less prominent source neglect for Mixed-lists (see Figure 3.4); however, the Bayesian evidence here was far too weak to draw conclusions, $BF_{10} = 0.73$.

One caveat with the comparisons between List membership and the present experiment, is that the data for the Source Similarity experiment are collapsed across conditions which are predicted to differ significantly. If one were to for instance compare List membership with the High-similarity (Within-modality) condition alone, there may be a greater difference between the two contexts.

Figure 3.4

Source Intrusion Monitoring by Output Position for List Membership and Mixed-List (Similarity) Contexts.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] List membership = 22 trials, Similarity = 44 trials

[b] List membership = 12 trials, Similarity = 18 trials

[c] List membership = 7 trials, Similarity = 16 trials

3.2.2.2 - *Source Similarity*

To test the hypothesis that there would be superior constrained search for the Low-similarity (Across-modality) lists compared with the High-similarity (Within-modality) lists, *t*-tests were conducted to observe if Similarity affected target and source intrusion availability. Minimum detectable effect size for .8 power was $d = 0.31$. There was no significant difference between the Similarity conditions in the number of targets generated, $t(63) = 0.56$, $p = .57$, $d = 0.07$, $BF_{10} = 0.16$, or the number of source intrusions generated $t(63) = 0.06$, $p = .96$, $d = 0.006$, $BF_{10} = 0.14$. The effect sizes here

are very small; therefore, target and source intrusion generation probably did not differ between the Similarity conditions. Collapsed across targets and source intrusions (PcSource), participants could selectively generate targets at above chance level (0.5) in the High-similarity condition, $t(63) = 6.08$, $p<.001$, $d = 0.76$, $BF_{10} = 3.45 \times 10^5$ and the Low-similarity condition, $t(63) = 6.87$, $p<.001$, $d = 0.86$, $BF_{10} = 6.91 \times 10^6$. In these cases Bayesian evidence for above chance performance was very strong. There was no significant difference in PcSource as a function of Similarity, $t(63) = 0.40$, $p = .34$, $d = 0.05$, $BF_{10} = 0.19$. Taken together, there is strong evidence from small effect sizes and conclusive Bayes Factors that Similarity did not affect constrained search. See Table 3.2 for descriptive statistics by Similarity.

Table 3.2

Targets and Source Intrusions Generated, Overall Search Accuracy and Proportions of Targets and Source Intrusions Monitored Correctly as a Function of Similarity.

Measure	High-similarity		Low-similarity	
	M	*SD*	*M*	*SD*
Targets	4.50	1.88	4.64	1.90
SI	2.43	1.67	2.42	1.53
PcSource	**0.65**	0.20	**0.66**	0.19
T mon (Pc)	**0.99**	0.04	**0.98**	0.05
SI mon (Pc)	**0.82****	0.28	**0.91****	0.20

Note. SI = Source intrusion, T mon = Target monitoring, SI mon = Source intrusion monitoring, M = Mean, SD = Standard Deviation. Bold font indicates significantly above chance performance.

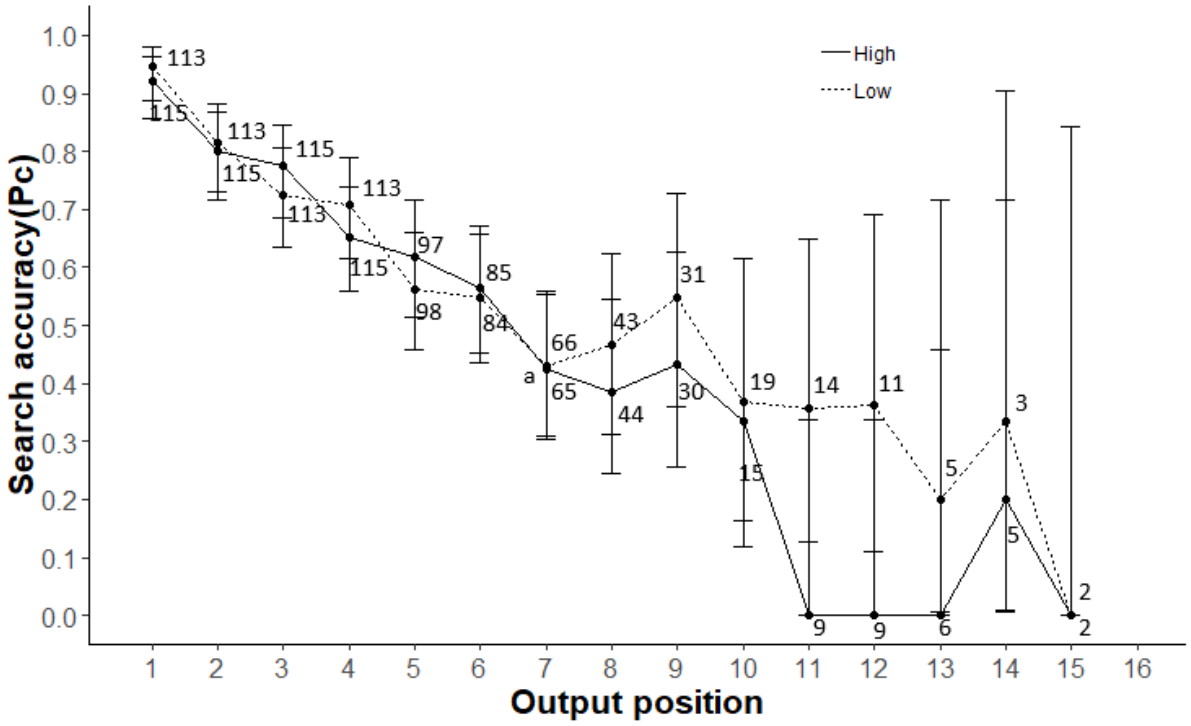
** $p < 0.01$

Search dynamics were examined to investigate whether there were any differences in search accuracy between the Similarity conditions at any stage in the recall period. BF_{10} for early (1-5), middle (6-10) and late (11-15) output positions were 0.06, 0.13 and 22.14 respectively. This implies that there is only an advantage for dissimilar sources compared to similar sources late on in the recall period. Further,

output position 1 was examined in isolation for a measure of retrieval cue distinctiveness unaffected by temporal context. BF_{10} for output position 1 was 0.11 indicating no difference between the Similarity conditions in cue distinctiveness. Search dynamics for the Similarity conditions are displayed in Figure 3.5.

Figure 3.5

Search Accuracy by Output Position for the High and Low-Similarity Conditions.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] High-similarity = 66 trials, Low-similarity = 65 trials

For both forms of monitoring, the minimum detectable effect size (assumed power = .8) for single sample and one-tailed paired t-tests was $d = 0.31$. Regarding target monitoring, single sample *t*-tests were conducted to observe if participants could monitor targets at above chance level. These revealed that participants were able to correctly accept targets at above chance level for the Within-modality trials, $t(63) = 108.94$, $p<.001$, $d = 13.62$, $BF_{10} = 5.78 \times 10^{69}$, and the Across-modality trials,

$t(63) = 70.31$, $p<.001$, $d = 8.79$, $BF_{10} = 1.18 \times 10^{58}$. Bayes Factors provide extremely strong support for above chance target monitoring in both Similarity conditions. One-tailed paired *t*-tests were conducted to observe if target monitoring accuracy differed as a function of Similarity. It was predicted that target monitoring would be superior for Across-modality lists than Within-modality lists. This was not found to be the case, $t(63) = -2.04$, $p = .98$, $d = 0.27$, supported by the Bayes Factor, $BF_{10} = 0.05$. See Table 3.2 for descriptives. This is probably not a power issue as the direction of the effect was the reverse of what was expected.

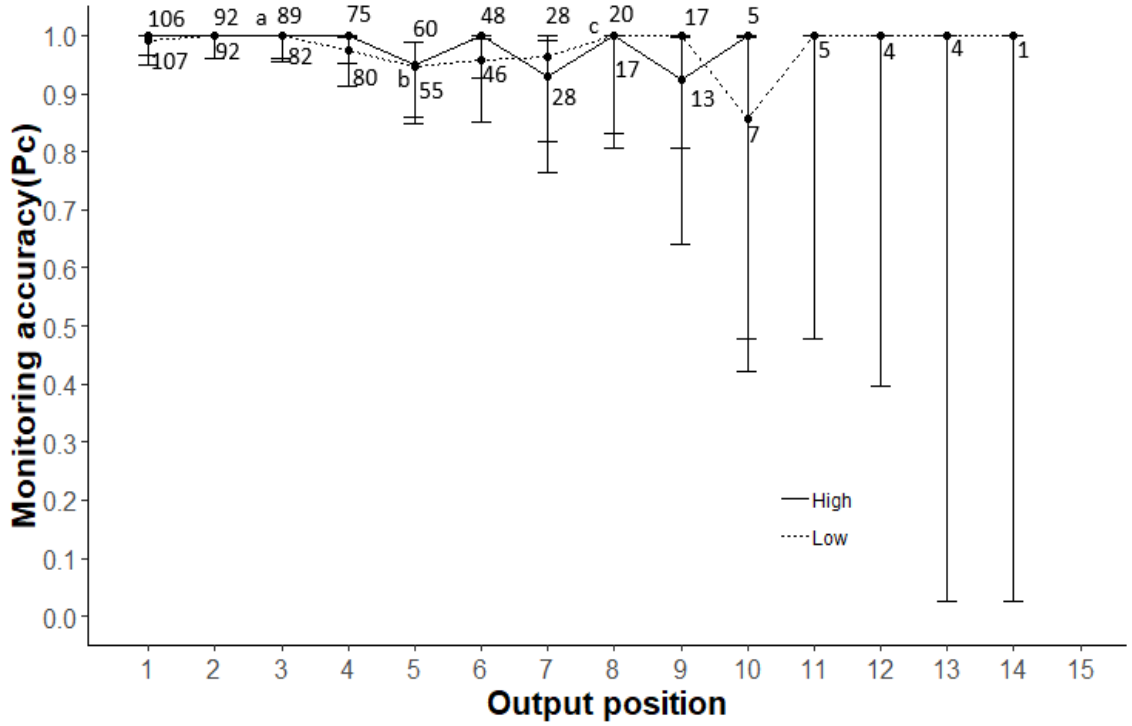
Target monitoring dynamics were then explored to investigate potential differences between the Similarity conditions at different stages in the recall period. The same analysis methods used for all previous recall dynamics analyses were applied. Only a single target was retrieved after output position 10 (14) for the High-similarity condition, so further output positions are disregarded. BF_{10} for early (1-3), middle (4-6) and late (7-10) output positions were 0.01, 0.10 and 0.10 respectively, implying no difference between the two Similarity conditions at any point in the recall period. See Figure 3.6 for target monitoring dynamics.

Identical analyses were conducted to assess source intrusion monitoring. Single sample *t*-tests revealed that participants were able to correctly reject source intrusions in the Within-modality lists, $t(63) = 9.12$, $p<.001$, $d = 1.14$, $BF_{10} = 4.03 \times 10^{10}$ and the Across-modality lists, $t(63) = 16.53$, $p<.001$, $d = 2.07$, $BF_{10} = 5.72 \times 10^{21}$. It was predicted that source intrusion monitoring would be significantly more accurate for the Across-modality lists than the Within-modality lists, due to Within-modality sources being more similar than Across-modality sources. A one tailed *t*-test revealed that this was indeed the case, $t(63) = 2.57$, $p = .006$, $d = 0.39$ supported by a Bayes Factor $BF_{10} = 5.60$. See Table 3.2 for descriptives. This demonstrates that source

intrusion monitoring was superior when sources were dissimilar compared to when they are similar.

Figure 3.6

Target Monitoring Accuracy by Output Position for the High and Low-Similarity Conditions.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] High-similarity = 89 trials, Low-similarity = 82 trials

[b] High-similarity = 60 trials, Low-similarity = 55 trials

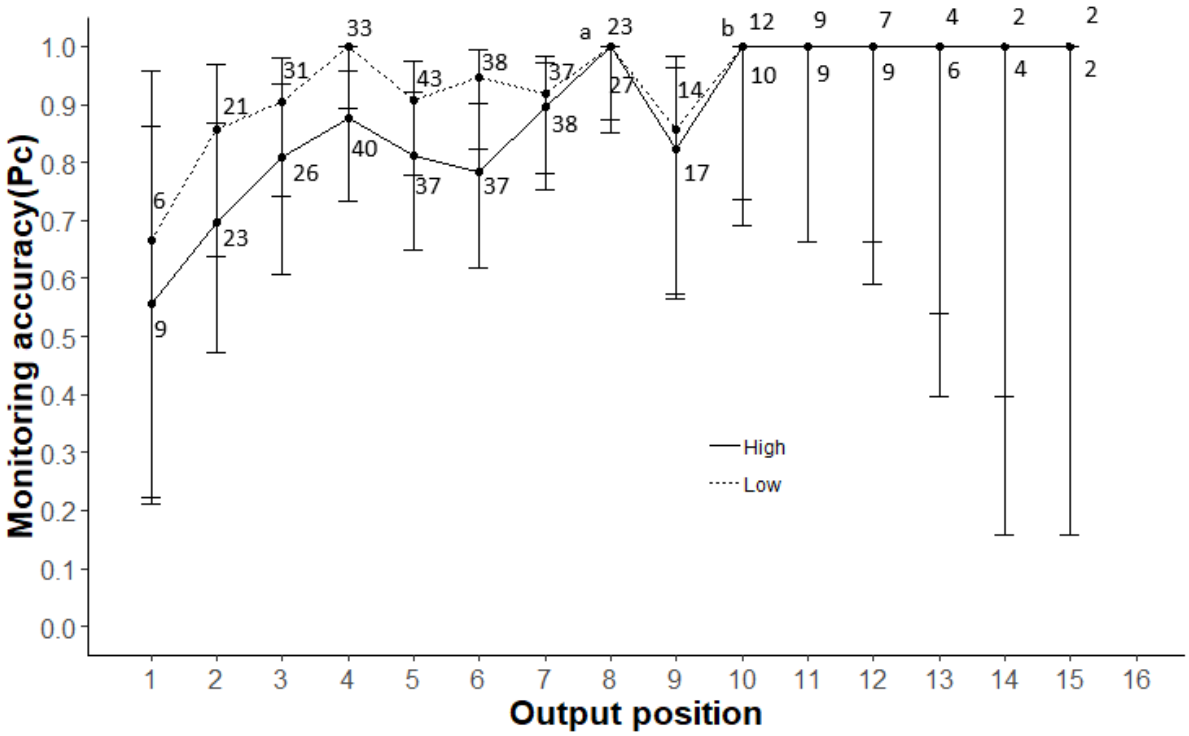
[c] High-similarity = 17 trials, Low-similarity = 20 trials

As with all previous output dynamics analyses, output positions were collapsed into early (1-5), middle (6-10) and late (11-15) for source intrusion monitoring. At early output positions there was evidence for a difference between the two Similarity conditions, BF_{10} = 4.35. At middle output positions, it would appear that there is no difference as a function of

Similarity however the evidence was weak, BF_{10} = 0.36. At late output positions, there was evidence that the Similarity conditions did not differ, BF_{10} = 0.07. Source intrusion monitoring dynamics are presented in Figure 3.7.

Figure 3.7

Source Intrusion Monitoring Accuracy by Output Position for High and Low-Similarity Conditions.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] High-similarity = 27 trials, Low-similarity = 23 trials

[b] For output positions 10-15, High-similarity = Below data point, Low-similarity = Above data point

3.2.3 - *Discussion*

The first aim of the present study was to assess the suitability of EFR as a method for simultaneously measuring constrained search and monitoring, given the concerns regarding selective reporting in EFR as discussed earlier. If it is a reliable method, then it was hypothesised that EFR should be sensitive to predictable

differences between List membership context as a source cue, and Mixed-lists. Namely that constrained search should be poorer for Mixed-lists than for List membership.

There is a tentative suggestion that constrained search accuracy is superior for List membership than for Mixed-lists, however the Bayesian evidence is too weak to make firm conclusions about this. In addition, taken separately there was no difference in target or source intrusion availability as a function of Context; however, the evidence was again far too weak to draw conclusions about constrained search based on these measures. Despite this, there was evidence for superior search accuracy for List membership at middle output positions, with no difference prior to this. This is consistent with temporal context having contrasting effects on search accuracy for List membership and Mixed-lists. Participants appear to be generating targets and source intrusions at the same rate for both contexts at early output positions. For List membership, temporal context will likely continue to activate targets more often than not from that point onwards owing to the contiguous nature of the target source. Participants will then terminate search when they believe they cannot retrieve any more targets. However, as targets deplete in Mixed-lists, temporal context is increasingly likely to activate source intrusions as the 2 sources are not temporally separated. In addition, once source intrusions have been retrieved, the retrieval cue will become trained to the wrong source, causing more source intrusions to be generated. It is unlikely that temporal context yields a more distinctive retrieval cue than Mixed-list contexts as search accuracy for the first retrieval did not differ between the two contexts.

The crossover in search accuracy between the two contexts around output positions 10 and 11, is likely due to targets depleting at different rates across contexts. The sudden drop in search accuracy for List membership indicates exhaustion of

targets from the search set, whereas some targets still remain for Mixed-lists. It seems that we can be fairly confident that EFR can detect predictable differences between contexts in search dynamics, if not raw numbers of targets and source intrusions, and output bound search accuracy.

Bayesian evidence suggested that there was no difference between the two contexts in source intrusion monitoring, although this evidence was weak at early output positions. The trend in Figure 3.4 does however suggest a small superiority for Mixed-lists for most of the recall period, particularly at output position 1. Although this was not statistically detected, it is possible that source neglect was less prevalent in Mixed-lists than for List membership for this same reason, given the visual difference (Figure 3.4) between the two contexts at output position 1. However a much larger sample size would be needed to detect any differences at the earliest output positions individually, without collapsing.

The second aim of the present study was to examine whether Source Similarity affects the ability to successfully constrain search. The hypothesis was that constrained search should be poorer in lists where two sources are more similar (Within-modality), compared to lists where sources are less similar (Across-modality). No significant evidence for this assertion was found for either targets and source intrusions generated, or search accuracy collapsed across all generated items. The reason for this may originate from the metacognitive control processes required to constrain search. In order to do this, participants first set a retrieval strategy and then specify and elaborate appropriate retrieval cues that will allow them to retrieve targets. These retrieval cues are not necessarily limited to source and could include any internal metacognitive knowledge or external information about the stimulus, which serve to make the sources more distinct. For instance a participant may have noticed subtle

differences in the regional accents of the male and female voices (Midlands vs West Country) at encoding, which could be incorporated into the retrieval cue. The idea of additional source information assisting with constrained search will be examined in more detail in Experiment 3.3.

This is supported by the search dynamics insofar as there was no difference at all between the two Similarity conditions for the first half of the recall period, in addition to a lack of difference at output position 1 in isolation, which indicates no difference in cue distinctiveness between the Similarity conditions. Although this is inconsistent with evidence for superior search accuracy for Low-similarity trials at late output positions. However, by this point in the recall period, the sample size had fallen dramatically as can be seen in Figure 3.5. There is a possibility that this difference is a chance occurrence, and a replication with a larger sample size would be required to see if this effect is genuine.

The final aim of the present experiment was to examine the effects of Source Similarity on source monitoring. The Source Monitoring Framework makes a simple prediction that monitoring judgments should be more accurate when the candidate sources are less similar. This prediction was clearly supported by the present data. Participants correctly rejected a significantly greater proportion of source intrusions in the Low-similarity lists than in the High-similarity trials, although there was no difference in target acceptance rates as confirmed by Bayesian evidence. Source intrusion monitoring dynamics revealed that differences between the Similarity conditions were likely to arise earlier in the recall period, when only perceptual source information is predicted to be available to the participant. Monitoring decisions involving cognitive operations performed on the item at encoding, such as 'what accent did the voice have?', take longer and often require retrieval of supporting

memories, which participants may not have had time to retrieve at early output positions. Given only perceptual source information, monitoring judgments will be more accurate in lists where the two sources are less similar than more similar.

To summarise, the present study, and comparisons with previous experiments provided interesting insights into the roles of constrained search and monitoring in recall accuracy, and demonstrated evidence for EFR as a reliable and viable method of measuring these 2 processes. Constrained search seems to be resistant to manipulations of Source Similarity, potentially due to metacognitive operations involved with setting an appropriate search strategy and locating retrieval cues. It is unlikely that the manipulation simply didn't work, as EFR has been shown to be capable of detecting differences in search accuracy where they arise. Plus, the manipulation was strong enough to elicit a predicted difference in source monitoring between the two Similarity conditions. It would be interesting to see if constrained search is resistant to other manipulations that source monitoring is sensitive to.
The next experiment will pilot various forms of source contexts which can be used for Experiment 3.3.

3.3 - Experiment 3.2

This experiment served as a pilot study for Experiment 3.3. Participants were tested for their ability to constrain search using a variety of Mixed-list contexts. If constrained search is successful, these contexts would be used in the next experiment which will investigate the use of task-irrelevant source information to assist in constraining search and monitoring. A starting point for this investigation were contexts which participants can successfully source monitor, namely Font Colour and Screen Location (Mulligan, 2004), Font Size, (Starns & Hicks, 2005) and Background (Doerksen & Shimamura, 2001). As previously stated, in the latter study source

monitoring was only at chance for words with no emotional valence. However, it is possible that the background manipulation used, the colour of a border surrounding the word, may not have been sufficiently memorable. For the present investigation, background will be defined as a full screen background image of a tiger or a landscape. These exemplars were chosen to be striking visually, with a multitude of distinguishing perceptual features such as colour, texture and perspective (Stenberg, 2006) which make them more memorable than a background distinction based on colour alone.

3.3.1 - *Methods*

3.3.1.1 - *Participants*

Forty additional participants took part in this study (13 Male, 27 Female, Mean age = 22.30, *SD* = 3.11). Participants were either Psychology undergraduates receiving compulsory course credit in order to pass a module, or paid members of the public receiving £4 for half an hour of laboratory time.

3.3.1.2 - *Design*

This experiment comprised one within-subjects factor, Context. Participants completed three trials each comprising a single list of twenty-four randomly allocated to be remembered words. The number of items per trial was increased as it was believed that the subsequent experiment would be far more challenging than previous ones, and may need more items to gain sufficient recall. Each trial was assigned a single source Context. This was counterbalanced by participant number so that each Context appeared an equal number of times across the experiment, and in each trial number to avoid order effects. Allocation of participant numbers to Context order was achieved prior to the experiment by means of random sampling without replacement. Memory was tested three times over the course of the experimental session (four including practice). Recall on each trial was implemented thirty seconds after

presentation of the list. Participants completed one practice trial beforehand comprising twelve words in the context which was not tested in the experimental trials.

3.3.1.3 - *Materials*

Stimuli were seventy-two nouns selected from the same pool as Experiment 3.1. This and subsequent experiments was programmed in PsychoPy software (Peirce et al. 2019). For the Colour trials, items were presented in the centre of the screen, in either red or blue letters; Arial font; size 0.25 (PsychoPy experimental settings); against a plain white background. For the Size trials, items were presented in the centre of the screen in either large or small letters (size 0.4 or 0.1 PsychoPy experimental settings); Arial font, and black text against a white background. For Location trials, items were presented at the top or bottom of the computer screen (0,0.5 or 0,-0.5 PsychoPy experimental settings) against a white background, in black Arial font; size 0.25 (PsychoPy experimental settings). For Background trials, items were presented in the centre of the screen, in black Arial font; size 0.25 (PsychoPy experimental settings, against a full screen background image of either a tiger or a landscape. To ensure that words were visible against these backgrounds, items were superimposed on a translucent white box (85% opacity) just large enough to encompass the word. Stimuli for the practice list were twelve randomly chosen verbal equivalents of the Snodgrass and Vanderwart (1980) pictures.

3.3.1.4 - *Procedure*

Study phase

Participants were first informed that they were going to study four lists of words (one per trial) (practice list included), and that within each list the presentation of each word would vary along a single dimension of either font colour, font size,

screen location or background image. For each trial, the participants were instructed to remember as many words as they could from the current list (forgetting previous lists), and how each word was presented, for a recall test at the end of the trial. They were also informed that presentation of words would vary along a different dimension for each trial. For example in trial 1, items could be printed in either red or blue; for trial 2 items could appear at the top or bottom of the screen and so on.

For each experimental list, participants were presented with twenty-four words on the computer screen one at a time in one of the four contexts (colour, size, location, background) detailed in the previous section. Each word had a presentation duration of five seconds, with a one second inter-stimulus interval (ISI). Presentation duration was increased to five seconds in anticipation for a far more challenging subsequent experiment. Following presentation of the twenty-fourth item the digit-sorting-distractor task then appeared for thirty seconds in black text; Arial font; size 0.25 PsychoPy experimental settings; white background).

Test phase

The test phase was largely similar to Experiments 2.3 and 3.1. However, there was no choice of source for the participants to make as the specific source that they were required to recall for EFR was randomised for each trial. Immediately following the numerical distractor, the participant received the on-screen EFR instruction which corresponded to the context they had studied. These were 'Recall the words printed in red/blue letters', 'Recall the words printed in big/small letters', 'Recall the words printed at the top/bottom of the screen' or 'Recall the words presented against a tiger/landscape'. Participants were only asked to recall one of the two sources on each trial. Participants then performed EFR using the tablet device as described in Chapter 1. As before, participants wrote their retrievals in the 'target' and 'other' boxes as

appropriate using a stylus, pressing next after each one. Previous retrievals were not visible to the participant. They pressed the finish button when they could not remember any more items from the correct source. The second recall attempt used in previous experiments was removed due to concerns with carry-over effects as previously highlighted. Upon pressing space on the keyboard, participants received the message "The next list is about to begin. Press space when ready" on the computer screen. Pressing space began the study phase of the following trial. After recall of the final trial, this message read "The experiment is now over. Thank you for your participation." For a schematic representation of the experimental procedure, see Figure 3.8.

3.3.1.5 - *Scoring*

Scoring was largely identical to Experiments 2.3 and 3.1. However, in the present experiment, participants studied more items per trial than for Experiment 2.3 (List membership). To make search metrics comparable between these two experiments, target and source intrusion availability needed to be adjusted for the number of presented items. To do this, target and source intrusion availability is expressed as the proportion of all presented targets and wrong-source items that were generated. The proportion of targets generated (PTarget) is expressed as in Equation 3.2.

$$PTarget = \frac{t}{T} \tag{3.2}$$

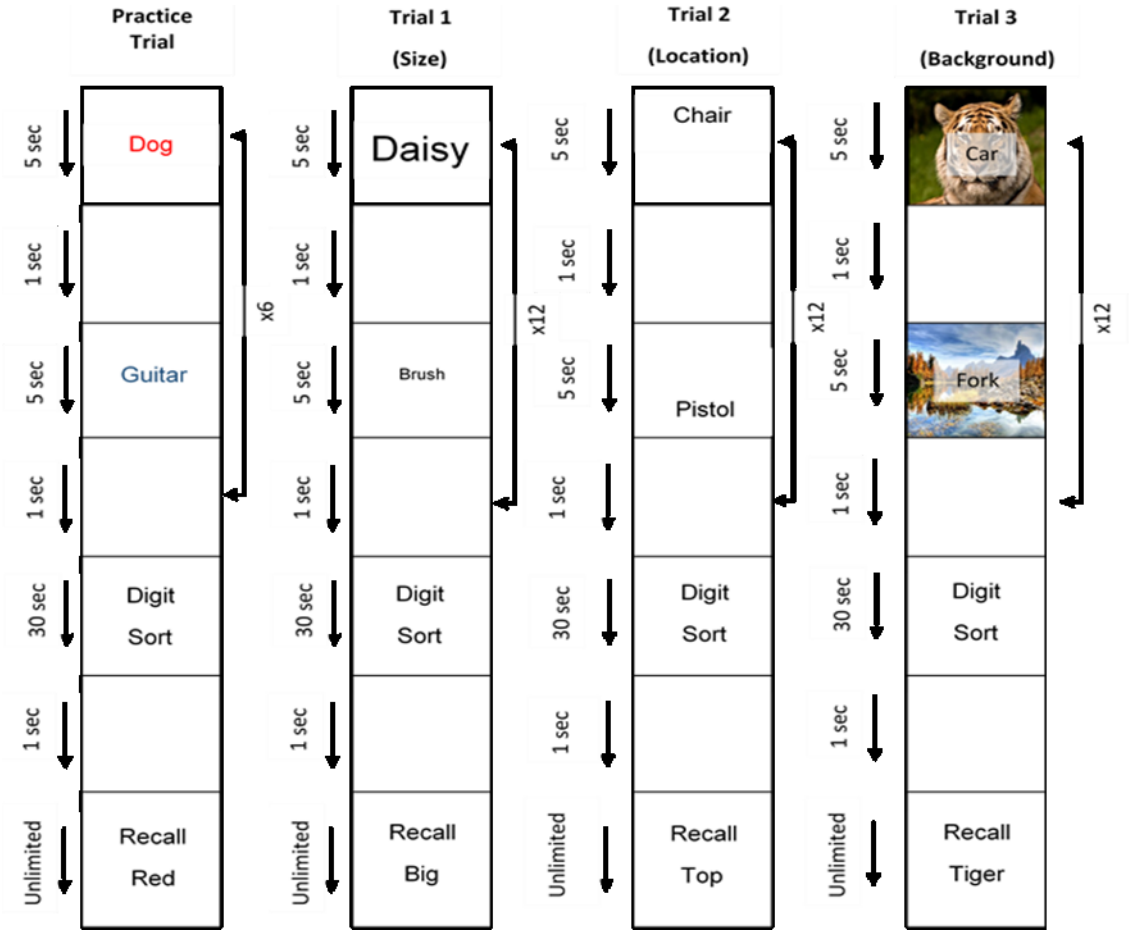
Where t is the number of targets generated by the participant and T is the total number of targets in the trial. Proportion of source intrusions generated (PSI) is expressed as in Equation 3.3.

$$PSI = \frac{s}{S} \tag{3.3}$$

Where s is the number of Source intrusions generated and S is the total number of presented wrong source items in the trial.

Figure 3.8

Schematic Representation of the Experimental Paradigm used for Experiment 3.2



Note. Digit sort = Digit sorting distractor task used throughout this thesis. This representation shows only one possible condition order. In reality any of these four contexts could be tested in any of the four trials according to counterbalancing. The precise source to be recalled on each trial was randomised.

3.3.2 – *Results*

A one-way (Context: Colour, Size, Location, Background) within-subjects ANOVA was conducted to observe if target availability differed between the 4 contexts. Minimum detectable effect size with assumed power of .8 for all ANOVAs reported in this experiment, was $\eta_p^2 = .03$. No significant difference was found, $F(3,117) = 1.89$, $p = .14$, $\eta_p^2 = .05$, supported by a Bayes Factor, $BF_{10} = 0.30$. A further ANOVA revealed that there was no significant difference in source intrusion availability among the four Contexts, $F(3,117) = 2.18$, $p = .09$, $\eta_p^2 = .05$. However, the Bayesian evidence was inconclusive, $BF_{10} = 0.44$.

Single sample *t*-tests (minimum detectable effect size assuming .8 power was $d = 0.40$) were conducted on PcSource scores for each of the contexts to observe whether participants could successfully constrain search using these contexts. Participants could successfully constrain search at above chance level (0.5) using Background context, $t(39) = 4.62$, $p<.001$, $d = 0.73$, $BF_{10} = 1079.56$, Colour context, $t(39) = 3.84$, $p<.001$, $d = 0.61$, $BF_{10} = 127.71$, and Size context, $t(39) = 6.09$, $p<.001$, $d = 0.96$, $BF_{10} = 8.28 \times 10^4$. Bayes Factors provide strong evidence for above chance performance in these three contexts. PcSource was not significantly greater than 0.5 for Location context, $t(39) = 1.63$, $p = .06$, $d = 0.26$; however, this may have been underpowered. In addition, Bayesian analysis revealed that there was almost no evidence for chance search accuracy, $BF_{10} = 1.08$. A one-way (Context: Colour, Size, Location, Background) within-subjects ANOVA revealed that there was a significant difference in PcSource across the four contexts, $F(3,117) = 2.91$, $p = .04$, $\eta_p^2 = .07$; however, the equivalent Bayesian analysis revealed that there was almost no evidence for this effect, $BF_{10} = 1.09$. P values for all pairwise comparisons were greater than the

Bonferroni corrected alpha value of .008. See Table 3.3 for descriptives. Bayesian pairwise comparisons are shown in Table 3.4. Posterior odds show that there was evidence for no difference in PcSource between Colour and Size, Colour and Location, Colour and Background, and Size and Background. Evidence for the null was extremely weak for the contrasts between Size and Location and Location and Background.

Table 3.3

Proportion of Targets and Source Intrusions Generated, Overall Search Accuracy, and Proportions of Targets and Source Intrusions Monitored Correctly Across Contexts.

Measure	Colour		Size		Location		Background	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
PTarget	0.44	0.16	0.50	0.13	0.44	0.21	0.47	0.20
PSI	0.31	0.16	0.31	0.14	0.36	0.18	0.29	0.15
PcSource	**0.59***	0.15	**0.63***	0.14	0.55*	0.18	**0.62***	0.16
T mon	**0.96**	0.12	**1.00**	0.02	**0.95**	0.17	**0.94**	0.11
SI mon	**0.81***	0.23	**0.92***	0.17	**0.83***	0.26	**0.91***	0.13

Note. T mon = Target monitoring, SI mon = Source intrusion monitoring, M = Mean, SD = Standard Deviation. Bold text indicates significantly greater than chance performance. For the significant effects of Context on PcSource and SI mon, it was not possible to determine which conditions significantly differed, as all *p* values for pairwise comparisons exceeded the Bonferroni corrected alpha value of *p* = .008.

* $p < 0.05$

Table 3.4

Bayesian Pairwise Comparisons for Overall Search Accuracy (PcSource) Scores Across Contexts

Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
Colour	Size	0.41	0.40	0.17
Colour	Location	0.41	0.66	0.27
Colour	Background	0.41	0.23	0.10
Size	Location	0.41	2.27	0.94
Size	Background	0.41	0.19	0.08
Location	Background	0.41	2.05	0.85

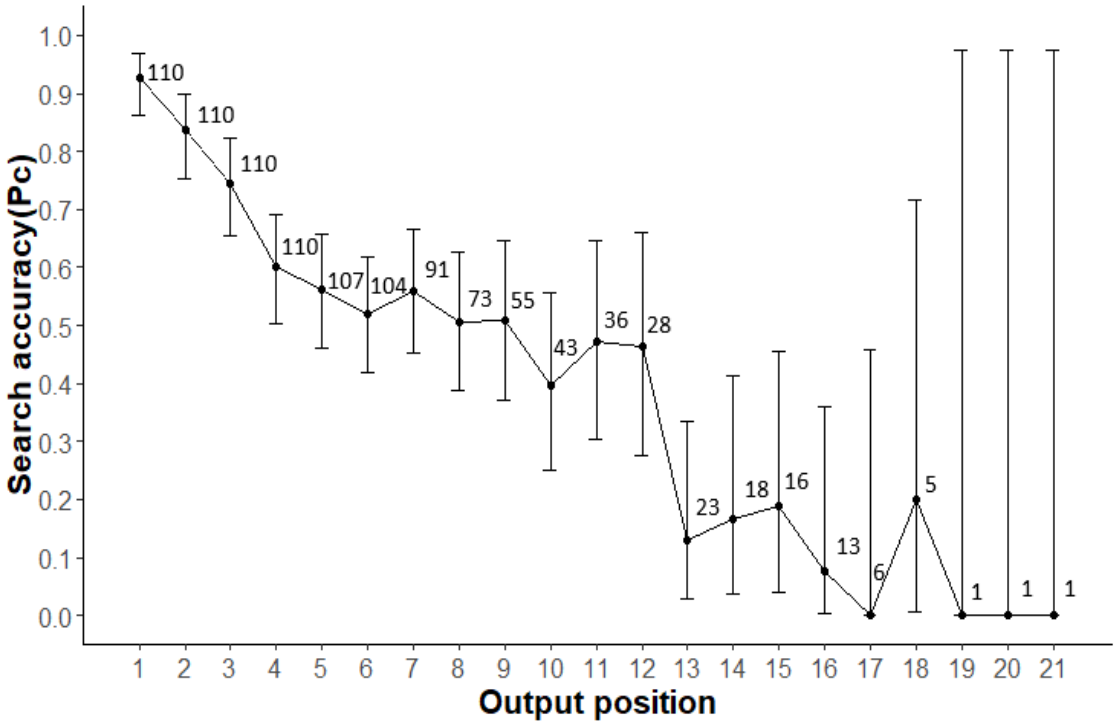
Search dynamics data partitioned into individual contexts was too noisy to perform effective contrasts; therefore, search dynamics are presented collapsed across contexts in Figure 3.9. The same general pattern as all previous EFR experiments is observed. Participants appear highly adept at retrieving targets at the beginning of the recall period, indicating successful context reinstatement, with a steady decline in search accuracy thereafter. See Chapter 5 for a discussion of the nature of this decline in search accuracy.

Single sample *t*-tests were conducted to observe if participants could monitor targets at above chance for each of the four contexts. Participants were able to successfully monitor targets at above chance level for Colour context, $t(39) = 24.44$, $p<.001$, $d = 3.86$, $BF_{10} = 2.11 \times 10^{22}$, Size context, $t(39) = 159.63$, $p<.001$, $d = 25.24$, $BF_{10} = 5.97 \times 10^{52}$, Location context, $t(39) = 16.87$, $p<.001$, $d = 2.67$ $BF_{10} = 5.80 \times 10^{16}$ and Background context, $t(39) = 25.29$, $p<.001$, $d = 4.00$, $BF_{10} = 7.22 \times 10^{22}$. All Bayes Factors indicate extremely strong evidence for these effects.

A one-way (Context: Colour, Size, Location, Background) within-subjects ANOVA was conducted to observe whether target monitoring accuracy differed as a function of Context. There was no significant difference found, $F(3,117) = 2.05$, $p = .11$, $\eta_p^2 = .05$; however, the Bayes Factor was inconclusive, $BF_{10} = 0.46$. As with search dynamics, monitoring dynamics for both targets and source intrusions was too noisy for individual contexts, so only the collapsed data is shown.

Figure 3.9

Search Accuracy by Output Position in Experiment 3.2.



Note. Data are collapsed across contexts, due to excessive noise within each context. Error bars represent 95% confidence intervals at each output position. Digits above each data point indicate the number of trials contributing data to that output position.

Further single sample *t*-tests were conducted to assess source intrusion monitoring accuracy in each context. Participants could successfully monitor source intrusions at above chance level for Colour context, $t(39) = 8.05$, $p<.001$, $d = 3.32$, $BF_{10} = 2.76 \times 10^{7}$, Size context, $t(39) = 15.56$, $p<.001$, $d = 5.42$, $BF_{10} = 4.04 \times 10^{15}$, Location context, $t(39) = 8.09$, $p<.001$, $d = 3.24$, $BF_{10} = 3.15 \times 10^{7}$ and Background context, $t(39) = 19.35$, $p<.001$, $d = 6.84$, $BF_{10} = 5.98 \times 10^{18}$. All Bayes Factors demonstrated extremely strong evidence for these effects. See Table 3.3 for monitoring descriptives.

A further one way within-subjects ANOVA was conducted to investigate whether there were differences between the contexts in source intrusion monitoring. This revealed that the proportion of source intrusions correctly rejected did differ as a function of Context, $F(3,117) = 3.17$, $p = 0.03$, $\eta_p^2 = .08$ although the evidence for this

effect was extremely weak, BF_{10} = 1.58. All p values from pairwise comparisons were greater than the Bonferroni adjusted alpha value of .008. See Table 3.5 for Bayesian pairwise comparisons. Posterior odds reveal evidence for no difference in source intrusion monitoring accuracy between Colour and Location, Size and Background and Location and Background. There was weak evidence for no difference between Colour and Background. Finally there was virtually no evidence for a difference, or lack thereof, between Colour and Size, and Size and Location.

Table 3.5

Bayesian Multiple Comparisons for Source Intrusion Monitoring Accuracy Across Contexts in Experiment 3.2.

Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
Colour	Size	0.41	2.44	1.01
Colour	Location	0.41	0.19	0.08
Colour	Background	0.41	1.37	0.57
Size	Location	0.41	2.27	0.94
Size	Background	0.41	0.18	0.07
Location	Background	0.41	0.67	0.28

Unfortunately there was too much noise in the monitoring dynamics for each individual context to perform comparisons; however, target and source intrusion monitoring dynamics were contrasted to gain insights into monitoring processes throughout the recall period. BF_{10} for early (1-5), middle (6-10) and late (11-16) output positions were 2.85 x 10^8, 1.14 and 18.74 respectively. No further output positions were analysed as only a single target was generated across all participants (output position 18) after this point. It appears that there is very strong evidence for superior target monitoring early in the recall period; however, we cannot conclude anything regarding middle output positions. There was evidence for a difference between target and source intrusion monitoring at late output positions. Unfortunately, we cannot

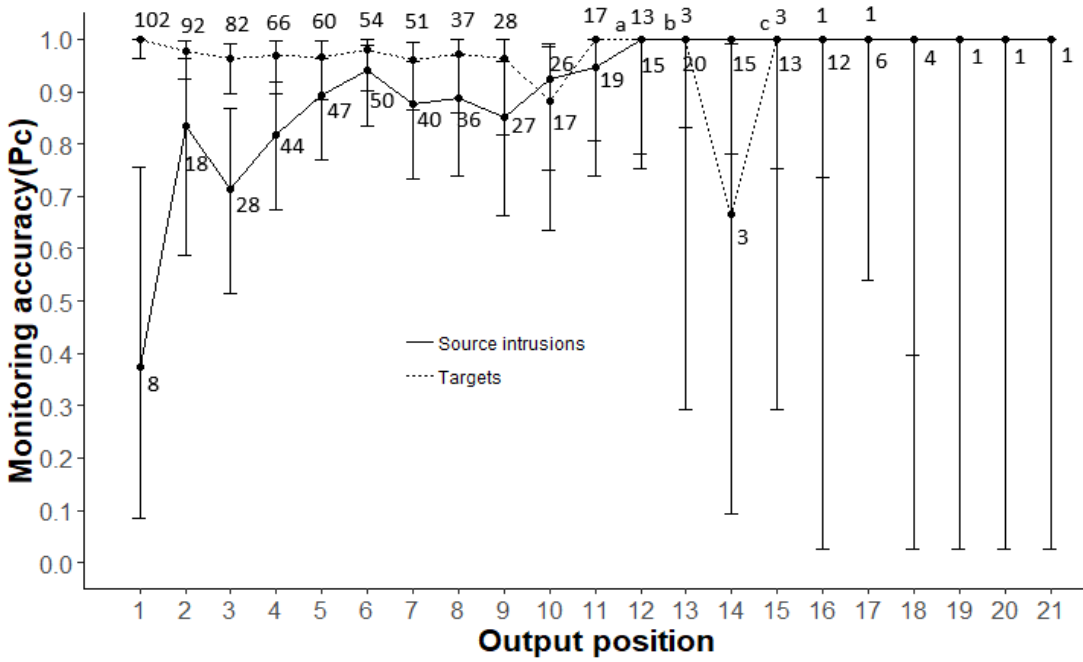
determine a direction as target monitoring accuracy is superior at output position 11, but source intrusion monitoring accuracy is superior at output position 14. As such, these output positions were examined in isolation to investigate where this difference might arise. There was evidence for no difference between the two types of monitoring at output position 11, $BF_{10} = 0.19$, however the evidence for a difference at output position 14 was too weak to draw conclusions, $BF_{10} = 1.78$.

A final point of interest was whether source neglect could be demonstrated by observing a large difference between target and source intrusion monitoring accuracy at output position 1. There was indeed strong evidence for superior target monitoring accuracy to source intrusion monitoring accuracy at this output position, $BF_{10} = 2.62 \times 10^5$, indicating strong evidence for source neglect.

The general trend again appears to show near ceiling target monitoring accuracy across most of the recall period. For source intrusion monitoring, the typical pattern of very poor performance at output position 1 followed by a progressive increase to ceiling thereafter was observed. See Figure 3.10 for target and source intrusion monitoring dynamics.

Figure 3.10

Target and Source Intrusion Monitoring by Output Position in Experiment 3.2



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] Targets = 13 trials, Source intrusions = 15 trials

[b] Targets = 3 trials, Source intrusions = 20 trials

[c] For output positions 15-21, Targets = above data point, Source intrusions = Below data point.

3.3.3 - *Discussion*

Upon examining Colour, Size, Location and Background contexts in isolation there was evidence that target availability, and weak evidence that source intrusion availability did not differ across the four contexts, which is largely promising. Analysis of constrained search collapsed across targets and source intrusions was mixed. There was evidence that participants could constrain search using Colour, Size and Background contexts. Constrained search for Location context was not significantly above chance; however, this analysis may have been underpowered, and the Bayesian

evidence was far too weak to confirm chance level performance. Similarly, the evidence for a significant difference in constrained search across contexts was ultimately too weak to attempt conclusions. Search dynamics collapsed across these contexts show a very similar pattern to Experiments 2.3 and 3.1. Very high search accuracy at output position 1 suggests successful reinstatement of source context, followed by a progressive decline in search accuracy thereafter, which will be explored further in Chapter 5.

Encouragingly, participants were able to monitor targets and source intrusions at above chance level in all four contexts, demonstrating that each context was sufficiently discriminable. There was a significant difference in source intrusion monitoring accuracy across contexts; however, again the Bayesian evidence was far too weak to draw a firm conclusion. Overall, from the Context comparisons conducted, it was deemed that all four contexts should be used in Experiment 3.3, as there was insufficient evidence that participants could not constrain search using any individual context. The inability to find significantly above chance search accuracy for location context may simply be an experimental power issue. In addition, the Bayes Factor was so close to 1, indicating no evidence at all for either chance or above chance performance.

Given that target monitoring accuracy was near ceiling for List membership, and for Mixed-lists in Experiment 3.1, it was unsurprising that this was the same for the present experiment collapsed across contexts. When source intrusion monitoring dynamics were examined, like List membership, there is very strong evidence for source neglect at output position 1, with participants seemingly demonstrating a strong bias to monitor this output position as a target. It is unlikely that participants were attempting to monitor output position 1 but were unable to do so, as target and

source intrusion monitoring would have been at chance if this were the case. From output position 2 onward, participants were appearing to engage in monitoring as the gap in performance between target and source intrusion monitoring greatly reduced. Late in the recall period there was a difference between the two types of monitoring; however, it was difficult to determine directionality. The pattern of source intrusion monitoring accuracy is in line with the Source Monitoring Framework (Johnson et al. 1993), which implies that source intrusion monitoring accuracy should improve as the recall progresses, owing to more source information being available as retrieval slows. The next experiment will examine whether participants can use task- irrelevant source information to assist with constraining search and monitoring.

3.4 - Experiment 3.3

In Experiment 3.1, it was found that constrained search was resistant to a manipulation of Source Similarity, which did affect source monitoring. A potential reason for this is expressed in more detail in section 3.2.6. However, in short, it is believed that participants may have been using non-experimentally manipulated source-identifying information to increase the distinctiveness of the two sources in the High-similarity (within-modality) condition. This additional source-identifying information is likely to consist of distinguishing features between the two voices noticed by participants during encoding, which could include differences in regional accent, pitch or tone. In Experiment 3.1 there was no attempt made to control for this, as it was not the purpose of the investigation. The present experiment aims to test the use of additional source information as an aid to constrain search and monitoring formally.

To do this, an experiment was designed whereby additional source information is either useful or not useful to the participant. Individual words are presented against

four different source contexts (Colour, Font size, Screen location and Background image, as in Experiment 3.2), in one of three Dependency conditions. In the Dependent source condition, the sources are completely predictive of one another (e.g. if the item is printed in red, the probability that it is presented at the top of the screen is 1). In the Independent source condition, the sources have no relationship, (e.g. if the item is red, then the probability that it is presented at the top of the screen is 0.5). The Partially dependent condition is between these two extremes, (e.g. if the item is red then the probability that it is presented at the top of the screen is 0.75). This arrangement of Dependency formalises the relationship between task-relevant and additional source-identifying information which may have played a role in Experiment 3.1. For example, in the High-similarity condition, the male voice having a Midlands accent, and the female voice having a West Country accent is akin to the Dependent source condition in the present experiment, as accent is completely predictive of voice gender.

As in Experiment 3.2, participants were asked to search (and monitor) for a particular subset of items cued by a source at test. It was expected that the availability of non-target source details (e.g. an item's location when the cue is to recall by colour), would help participants identify target items most when sources were predictive of one another, but not when they were Independent, with Partially dependent sources falling between these two cases. What was of particular interest was whether the relationship between sources would have their effects on search accuracy or monitoring. If the relationship between sources aids search, then the highest proportion of targets recalled in output position 1 should be in the Dependent condition, followed by the Partially dependent and Independent conditions respectively. Monitoring should show a similar pattern of results. In the Dependent

condition monitoring judgments can be made upon comparisons between the retrieved item and a greater number of retrieval cues. Therefore, inferences can be made to assist in source monitoring. In the Independent condition, judgments can only be made between the source of the retrieved item and a single context cue. Therefore, source confusion is more likely as there is less source information for the participant to draw upon. The Partially dependent condition should yield intermediate source monitoring performance.

An alternative explanation is that the search only proceeds on the basis of a single cue, and that the benefits of Source Dependency emerges during retrieval monitoring. Maylor et al. (2001) asked participants to generate items from either a single semantic category (e.g. foods, countries), or both of these categories simultaneously (foods or countries). It was found that the rate of retrieval for the joint category condition was no faster than the quicker of the two single categories, indicating that participants could not search for two categories in parallel. The same result was found for long-term autobiographical memory, where participants were required to retrieve autobiographical memories associated with particular cue words e.g. (flower, ticket). If memory search can only progress based on a single cue then there will be no constrained search advantage for the Dependent condition relative to the Partially dependent or Independent conditions. Although monitoring may still benefit from full Source Dependency in the manner previously described.

3.4.1 - *Methods*

3.4.1.1 - *Participants*

Forty participants were recruited for this study (14 Male, 26 Female, Mean age = 21.03, *SD* = 3.53). Participants were either Psychology undergraduates requiring

compulsory course credit in order to pass a module or paid members of the public receiving £4 for half an hour of participation time.

3.4.1.2 - *Design*

There was one within-subjects factor, Context Dependency, comprising three levels. In the Dependent condition, all source contexts directly related to each other. For example all red items appeared in a large font, at the top of the screen and against the background of a tiger. Therefore the probability of Dependency between any two source contexts was 1. In the Partially dependent condition, the mean Dependency between any two source contexts was p=0.75. In a third, Independent condition, each item was randomly assigned one source from each context, yielding a stochastic probability of 0.5. Seventy-two concrete nouns were randomly allocated to one of the three trials. Condition order was counterbalanced by participant number to avoid order effects. Contexts were also counterbalanced by participant number so that each context was tested an equal number of times for each condition across the whole experiment. Allocation of participant numbers to a Dependency condition order and context order was conducted prior to the experiment by means of random sampling without replacement. Memory was tested three times over the experimental session (four with practice trial). The recall test for each list was implemented immediately following the digit-sorting-distractor task of the study phase.

3.4.1.3 - *Materials*

Stimuli were identical to Experiment 3.2. Each item was presented in Arial font, and in four sources simultaneously: one of two font colours, one of two screen locations, one of two font sizes and one of two background images (see section 3.3.1.3). All items were presented against a translucent white box (85% opacity) large enough to encompass items printed in the larger font size. Allocation of sources to

items was determined according to Dependency condition, as described in section 3.4.1.2.

3.4.1.4 - *Procedure*

Study phase -

Participants were informed that they would see a list of words on the screen, and that they should try and remember as many words as they could from the present list, as well as how each word was presented. Participants completed one practice list of twelve words to familiarise them with the procedure followed by three experimental lists. Stimulus presentation timings were identical to Experiment 3.2 (see section 3.3.1.4). At the end of each list, participants completed the same thirty-second digit-sorting-distractor task as all EFR experiments in this thesis.

Test phase -

Following the numerical distractor, a screen appeared prompting the participants to 'Choose A, B, C or D'. They were informed that this choice would determine which of the four study contexts they should use as a cue for recall. They were also told that for each list, the choice would be randomised, so any context could correspond to any of the four keys. This ensured that participants could not expect a particular context cue for any given list, and would need to pay attention to all four study contexts at study. In reality the order of context cues for each participant across lists was counterbalanced by subject, so that each context was used as a recall cue the same number of times across the whole experiment. Upon pressing one of the four keys, a screen appeared with a recall instruction relating to one of the four contexts. These were: Recall the items printed in red/blue, recall the items printed in big/small letters, recall the items printed at the top/bottom of the screen, and recall the words presented against a tiger/landscape. For each context cue, the specific source that

participants were required to recall was randomised so there was an equal chance of recalling either source for each context.

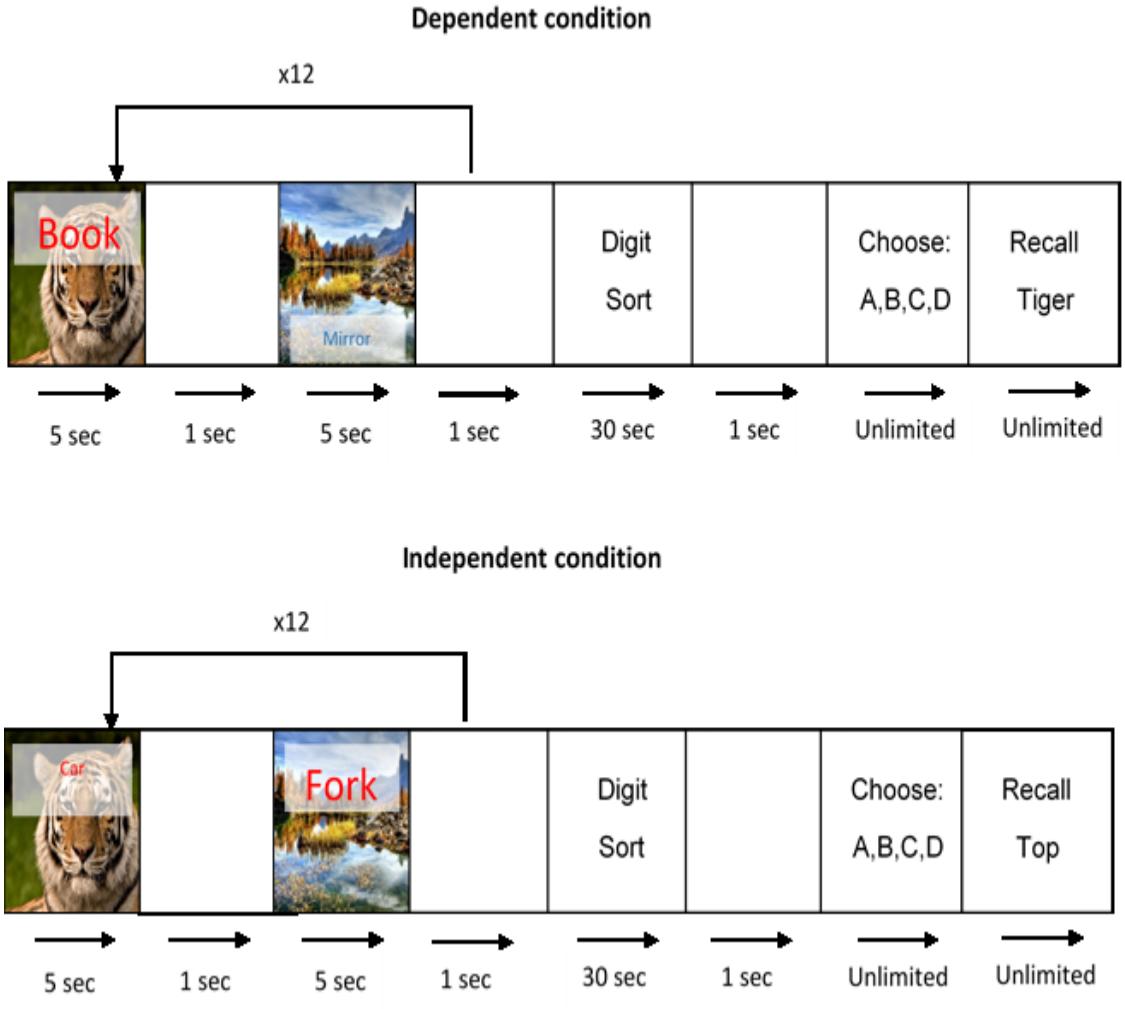
The tablet device was handed to the participants at this point for the recall test. EFR instructions were identical to all previous EFR experiments. Participants wrote their retrievals using a stylus in the 'target' or 'other' box as appropriate, clearing the screen with the 'next' button in the top-right corner of each box. Previously recalled items were not visible to participants once 'next' had been pressed. As with the previous experiment there was no second recall attempt due to concerns with carry-over effects. Once participants had finished recalling items, they pressed the 'finish' button on the tablet. The tablet was then returned to the experimenter. Upon pressing space on the computer keyboard, one of two messages was displayed on the computer screen. On experimental trials 1 and 2 the message was 'The next list is about to begin. Press <space> when ready'. Pressing space began the study phase of the next list. After the final experimental trial the message read 'The experiment is now over. Thank you for your participation'. See Figure 3.11 for a schematic representation of the paradigm.

3.4.1.5 - *Scoring*

Scoring was identical to Experiment 3.2.

Figure 3.11

Schematic Representation of the Experimental Paradigm for Experiment 3.3.



Note. Digit sort = Digit sorting distractor task used throughout this thesis. For the Dependent condition there were two possible combinations of sources for each item as demonstrated. For the Independent condition each item was randomly allocated a source for each context. The Partially dependent condition has been excluded as the screen appearance was largely the same as the Independent condition. The difference was that allocation of sources to items was not entirely random, as described in section 3.4.1.2.

3.4.2 - *Results*

A one-way (Dependency: Dependent, Partially dependent, Independent) within-subjects ANOVA was conducted to investigate differences in target availability between the Dependency conditions. For all ANOVAs reported in this experiment, the minimum detectable effect size for assumed power .8 was $\eta_p^2 = .04$ There was no significant effect of Dependency on proportion of targets generated (PTarget), $F(2,78)$

= 1.63, p = .20, η_p^2 = .04, BF_{10} = 0.19. Another one-way (Dependency: Dependent, Partially dependent, Independent) within-subjects ANOVA revealed that there was no significant effect of Dependency on proportion of source intrusions generated (PSIntrusion), $F(2,78)$ = 1.16, p = .32, η_p^2 = .03, BF_{10} = 0.17. This may have been due to low power, but Bayes Factors demonstrate credible evidence that the Dependency conditions did not differ in target or source intrusion availability.

Collapsed across targets and source intrusions, single sample *t*-tests (minimum detectable effect size was d = 0.40) were conducted to see whether participants could successfully constrain search at above chance level (PcSource = 0.5) in each of the three Dependency conditions. Participants could not successfully constrain search at above chance level in the Dependent condition, $t(39)$ = 0.26, p = .40, d = 0.04, BF_{10} = 0.21, the Partially dependent condition, $t(39)$ = -1.28, p = .90, d = 0.20, BF_{10} = 0.08, or the Independent condition, $t(39)$ = 1.17, p = .13, d = 0.18, BF_{10} = 0.56. Although there is a suspicion of a lack of power in this analysis, Bayes Factors indicate that there was credible evidence for a lack of ability to constrain search in the Dependent and Partially dependent conditions, but not in the Independent condition. A one-way (Dependency: Dependent, Partially dependent, Independent) within-subjects ANOVA revealed no significant effect of Dependency on PcSource scores, $F(2,78)$ = 1.62, p = .20, η_p^2 = .04, supported by a Bayes Factor, BF_{10} = 0.30. See Table 3.6 for descriptive statistics of search metrics.

Table 3.6

Proportion of Targets and Source Intrusions Generated, Overall Search Accuracy and Proportions of Targets and Source Intrusions Monitored Correctly for each Dependency Condition.

Measure	Dependent		Partially dependent		Independent	
	M	*SD*	*M*	*SD*	*M*	*SD*
PTarget	0.31	0.16	0.26	0.15	0.31	0.18
PSI	0.29	0.14	0.31	0.14	0.26	0.14
PcSource	0.51	0.18	0.46	0.19	0.54	0.24
T mon	**0.90**	0.16	**0.90**	0.16	**0.89**	0.16
SI mon	0.58	0.35	0.51	0.29	**0.60**	0.28

Note. T mon = Target monitoring, SI mon = Source intrusion monitoring. M = Mean, SD = Standard Deviation. Bold text indicates significantly above chance performance.

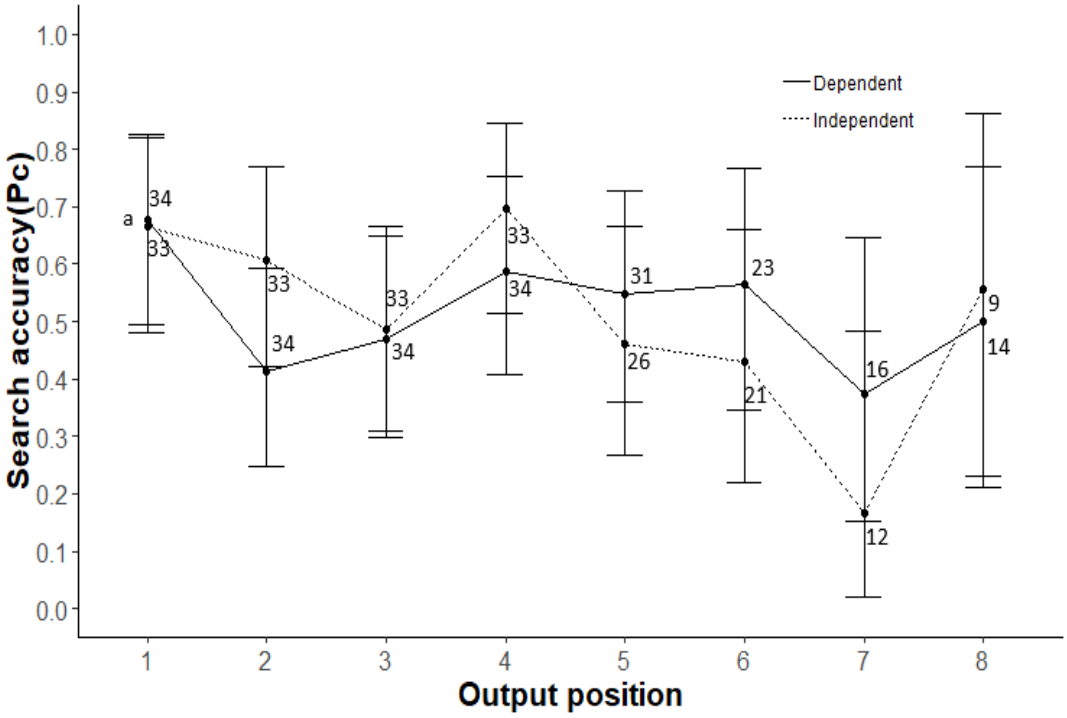
Search dynamics were investigated to detect any differences between the Dependency conditions at different stages during the recall period. BF_{10} for early (1-3), middle (4-6) and late (7-8) output positions were 0.10, 0.07 and 0.13 respectively, indicating evidence for no difference between the Dependency conditions throughout the recall period. A further important point of interest was whether there was an accuracy advantage for the Dependent condition at output position 1. This would indicate that participants were able to reinstate more than one context for constraining search. BF_{10} for output position 1 was 0.19, respectively indicating evidence for no difference between the conditions. In all conditions, the data do not follow the pattern seen in previous EFR experiments where search accuracy is very high at the beginning of the recall period, progressively falling with output position. Instead, performance appears to fluctuate across the recall period with no obvious trend (see Figure 3.12), which is consistent with search accuracy which does not exceed chance across output positions i.e. PcSource of 0.5.

Single sample *t*-tests were conducted to observe if participants could successfully monitor targets at above chance level in the three Dependency conditions. These revealed successful above chance target monitoring for the Dependent condition, $t(39) = 15.71$, $p<.001$, $d = 2.48$, BF = 5.49×10^{15}, the Partially dependent condition, $t(39) = 15.83$, $p<.001$, $d = 2.50$, $BF_{10} = 7.07 \times 10^{15}$, and the Independent condition, $t(39) = 15.21$, $p<.001$, $d = 2.40$, $BF_{10} = 1.91 \times 10^{15}$. Bayes Factors demonstrate extremely strong evidence for above chance target monitoring in all Dependency conditions. A one-way (Dependency: Dependent, Partially dependent, Independent) within-subjects ANOVA revealed no significant effect of Dependency on target monitoring accuracy, $F(2,78) = 0.09$, $p = .92$, $\eta_p^2 = .002$, supported by a Bayes Factor, $BF_{10} = 0.08$. See Table 3.5 for descriptive statistics of target monitoring accuracy.

Target monitoring dynamics were examined using Bayesian contingency tables to investigate potential differences between the Dependency conditions at different stages of the recall period. BF_{10} for early (1-3), middle (4-6) and late (7-8) output positions were 0.11, 0.04 and 0.55 respectively Overall the Bayes Factors suggest evidence for no difference in target monitoring between the Dependency conditions early and mid-way through the recall period. However, there is not enough evidence to make conclusions about late output positions. See Figure 3.13 for target monitoring dynamics.

Figure 3.12

Search Accuracy by Output Position as a Function of Dependency.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position. Partially dependent condition has been removed for clarity.

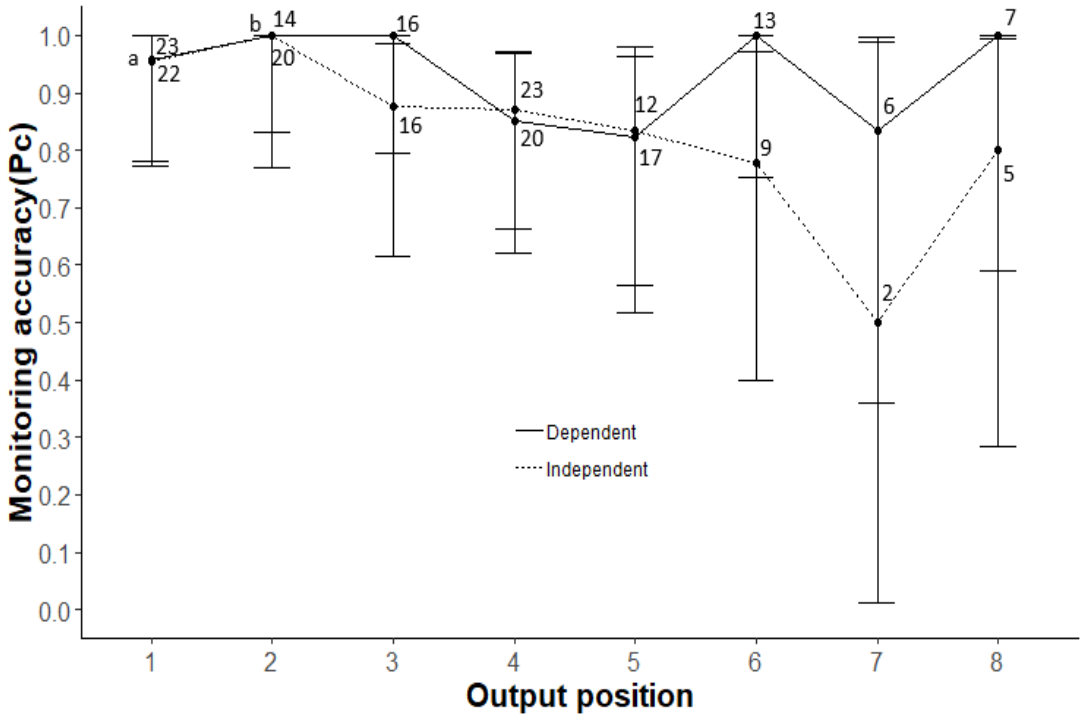
[a] Dependent condition = 34 trials, Independent condition = 33 trials

Single sample *t*-tests were conducted to observe if participants could successfully monitor source intrusions at above chance level in each of the Dependency conditions. Participants were unable to monitor source intrusions at above chance level in the Dependent condition, $t(39) = 1.47$, $p = .07$, $d = 0.23$, however the Bayes factor demonstrated that evidence for this was extremely weak, $BF_{10} = 0.85$, and there may be a suspicion of low power in the analysis. Participants were unable to monitor source intrusions at above chance level in the Partially dependent condition, $t(39) = 0.11$, $p = .46$, $d = 0.02$ supported by the Bayes Factor $BF_{10} = 0.19$. However participants were able to monitor source intrusions at above chance level in the Independent condition,

$t(39) = 2.20$, $p = .02$, $d = 0.35$. However, this effect size is lower than the minimum detectable effect size, implying that one can still happen to observe this significant difference, although less than four times out of five when the alternate hypothesis is true as dictated by power of .8. This is reinforced somewhat by an inconclusive Bayes Factor, $BF_{10} = 2.88$. Therefore, we cannot draw firm conclusions as to whether participants can truly monitor source intrusions at above chance level in the Independent condition.

Figure 3.13

Target Monitoring Accuracy by Output Position as a Function of Dependency



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position. Partially dependent condition has been removed for clarity.

[a] Dependent condition = 23 trials, Independent condition = 22 trials

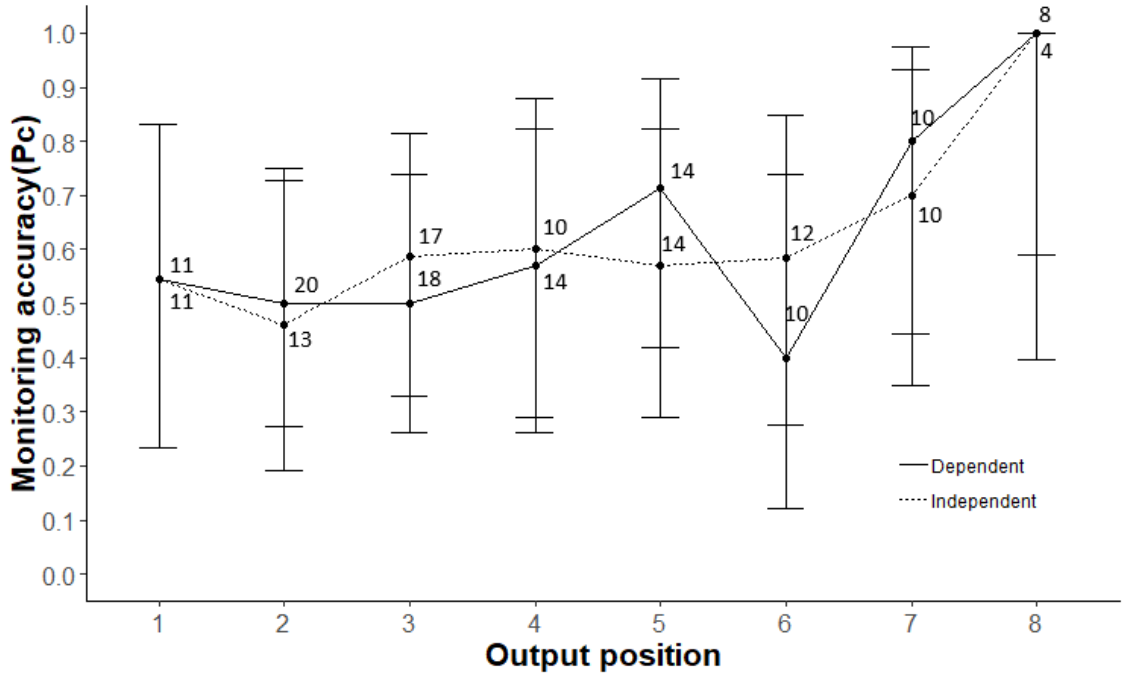
[b] Dependent condition = 14 trials, Independent condition = 20 trials

A one-way (Dependency: Dependent, Partially dependent, Independent) within-subjects ANOVA revealed that there was no significant effect of Dependency on intrusion monitoring accuracy, $F(2,78) = 1.14$, $p = .33$, $\eta_p^2 = .03$. Although this analysis may have been underpowered, the Bayesian analysis demonstrated credible evidence that the Dependency conditions did not differ in Source intrusion monitoring accuracy, $BF_{10} = 0.18$.

The same analysis conducted on target monitoring dynamics was also applied to source intrusion monitoring dynamics to detect potential differences as a function of Dependency for each output position. BF_{10} for early (1-3) middle (4-6) and late (7-8) output positions were 0.38, 0.11 and 0.14 respectively. This implies that there was no difference in source intrusion monitoring between the Dependency conditions at middle and late output positions. However, there was not enough evidence to draw conclusions regarding early output positions. One notable difference between source intrusion dynamics for this experiment and previous EFR studies presented in this thesis, is the lack of improvement in accuracy from output positions 1-2. This is true of all Dependency conditions. In fact, for all three conditions there is a surprising drop in monitoring accuracy from output positions 1-2. See Figure 3.14 for source intrusion monitoring dynamics (Partially dependent condition not shown).

Figure 3.14

Source Intrusion Monitoring Accuracy by Output Position as a Function of Dependency.



Note. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position. Partially dependent condition has been removed for clarity.

3.4.3 - *Discussion*

This experiment sought to investigate whether participants can use additional source information to assist with constraining search. It was predicted that there should be a relationship between how predictive this additional source information is of the target source, and search and monitoring accuracy. Performance should be highest when the additional source information is totally predictive of the target source (Dependent condition), followed by when it is partially predictive (Partially dependent condition). Poorest performance should be in a condition where the additional source information does not predict the target source (Independent condition). An alternative is that search can only be accomplished using a single source

cue, in other words participants cannot search for more than one source at a time. If this is the case then there should be significantly above chance search performance in all Dependency conditions, but no difference as a function of Dependency.

Evidence from various analyses points to there being little evidence for either of these assertions. Context Dependency had no effect on target and source intrusion availability, or constrained search accuracy collapsed across item type (PcSource). Search dynamics largely supported this, as there was evidence for no effect of Dependency throughout the recall period. This should not be taken as support for exclusively searching for one source at a time, as participants could not constrain search at above chance level in any of the Dependency conditions. The fact that there was no trend of decreasing search accuracy over time, irrespective of Dependency condition, suggests that the pool of targets in the search set was not being depleted very quickly. This is consistent with participants having great difficulty in preferentially retrieving targets over incorrect-source items. CMR provides a plausible explanation for the lack of constrained search, in that the source context cue may have been too weak to overcome temporal or semantic associations between items, possibly due to the difficulty of simultaneously encoding four source contexts. Therefore, incorrect items may have received a similar degree of support from the retrieval cue as targets.

Regarding monitoring, again there is little support for the prediction that source Dependency should aid monitoring. Dependency had no significant effect on target acceptance or source intrusion rejection rates. When we examine monitoring dynamics, this appears to tell the same story. At most output positions there was evidence for no difference in target or source intrusion monitoring accuracy between the Dependency conditions. In fact, source intrusion rejection rates demonstrate that participants appeared to have great difficulty monitoring source intrusions. There was

only weak evidence that they were able to monitor source intrusions in the Independent condition. This is further evidence of poor source encoding.

3.5 - General discussion

The present chapter aimed to build on Chapter 2, by exploring constrained search and monitoring in situations where participants cannot use temporal associations among items to assist in retrieving targets. This offered an opportunity to test the reliability and viability of EFR as a paradigm for simultaneously measuring constrained search and source monitoring. One possible concern with EFR is the potential for selective reporting, where participants only report items that they believe are correct; thus, reducing the sensitivity of the paradigm to detect the effects of a manipulation on item generation. If EFR is indeed a viable methodology, then it should be sensitive enough to detect predictable differences in constrained search. It was hypothesised that constrained search should be poorer in Mixed-lists compared with List membership as presented in Chapter 2. This is owing to the removal of confounding temporal factors which may have aided constrained search in temporally separated sources. As such, one would expect a significant reduction in constrained search accuracy in Mixed-list experiments if participants are not selectively reporting correct items.

Evidence from Experiment 3.1 tentatively suggests that selective reporting is not a significant confound, at least in the context of this thesis. Firstly, search accuracy collapsed across targets and source intrusions was significantly less for Experiment 3.1 than List membership, although not at .8 power. In the search dynamics data this manifested in superior search accuracy in middle output positions (5-8).

Search on the whole appeared to be quite resistant to source manipulations within experiments. For example it was predicted that search accuracy would be

significantly poorer when the sources to be distinguished were more similar than less similar. In fact, there was evidence for no effect of Source Similarity on target availability, source intrusion availability or overall search accuracy. The only difference noted was superior search accuracy for Low-similarity at late output positions; however, this should be treated with caution, given that few participants contributed to these output positions. One potential account for this pattern was that participants noticed additional source identifying information regarding the two voices in the High-similarity condition during encoding. They were then able to use this information to more effectively search for targets.

To test this formally, an attempt was made to manipulate constrained search by making additional contextual information available to the participant at retrieval. Although on the surface it could be interpreted that making additional contextual information available to the participant does not aid constrained search, one must bear in mind that participants could scarcely constrain search at all in this experiment. This may be due to the fact that each item was associated with four sources as opposed to one for all previous experiments, which could have led to poorer encoding of source due to the need to remember the item and all sources associated with it. This would certainly account for poor search accuracy in all Dependency conditions. Another unfortunate effect of task difficulty could be that participants might not have noticed the Dependency manipulation, which would explain the lack of an effect of Dependency. Thus, it would be premature to state that Context Dependency truly has no effect on constrained search. Future experiments may wish to reduce the number of source contexts utilised in order to make the task less taxing.

For source monitoring, similar patterns of results to List membership were obtained. Consistently high target monitoring accuracy was observed in Experiments

3.1-3.3. Although it was predicted that source intrusion monitoring should be superior for Mixed-lists than for List membership, this was not found. The origin of these effects would most likely be at the beginning of the recall period, where participants only have access to perceptual source information for monitoring judgments. Unfortunately at early output positions very few source intrusions are generated, so this predicted difference may have been hard to detect. There is a suggestion of this in the trend for monitoring dynamics in Figure 3.4. Although there is not Bayesian evidence to support this due to the sparse data at output position 1, it would appear that there is numerically superior source intrusion monitoring for Mixed-lists than List membership, which is indicative of reduced source neglect.

Source monitoring did show the predicted effects of Source Similarity in Experiment 3.1. Participants rejected a significantly greater proportion of source intrusions correctly in the Low-similarity condition than the High-similarity condition. Monitoring dynamics suggested that this advantage for Low-similarity occurred early in the recall period where retrieval is assumed to be rapid, and according to the Source Monitoring Framework (Johnson et al. 1993) perceptual source information is most integral to monitoring judgments.

Unfortunately, conclusions could not be drawn regarding the potential effects of Context Dependency on source monitoring, owing to participants having great difficulty monitoring source intrusions. This is likely to be for the same reason as chance constrained search performance in the same experiment. The requirement to simultaneously study four sources for each item led to poor source encoding, and consequently poor source intrusion monitoring. The next chapter will examine an alternative method of measuring constrained search which is not affected by selective reporting confounds.

Chapter 4: Search set size estimation

4.1 - Introduction

So far, this thesis has utilised Externalised-Free Recall (EFR) to examine an individual's ability to constrain their search by source, and to monitor the output of that search to achieve accurate output. This procedure provides a rich examination of what a participant has searched, and the simultaneous success of monitoring for each retrieval. However, as alluded to previously in this thesis, it is potentially susceptible to selective reporting. Experiments in Chapter 3 demonstrate that EFR can detect predictable differences in constrained search between different forms of context, suggesting that selective reporting is not completely confounding constrained search measures. However, this does not mean that selective reporting is absent, and could not still be overestimating constrained search performance. Therefore, it is important to compare EFR with other methodologies where selective reporting is presumably absent.

A viable option is examination of recall latencies, the precise timings of overt retrievals during recall. In the 1950s Bousfield and colleagues (Bousfield et al., 1956, 1958) conducted multiple experiments which investigated the relationship between item output order and probability of recall. The general finding across these studies was that items which have the highest probability of being recalled, have a tendency to be recalled before items with lower recall probabilities. That is, items with a high probability of recall will have relatively shorter recall latencies compared with those which have a low probability of recall.

Indeed this is a prediction made by retrieved context models such as CMR (Polyn et al. 2009a) and CMR2 (Lohnas et al. 2015). In these models, an item's match with the current state of context is directly related to its degree of support in an ensuing retrieval competition. Items with the greatest match to current context receive the most support and are the most likely to be retrieved. The items which most strongly match the current state of context at the start of the recall period are recency items. Items with the poorest match with the current state of context at the start of the recall period may not be included in the search set. An exception is made for primacy items. Primacy is modelled as increased attention afforded to the earliest items in the list; therefore, they have stronger memory traces. With each recall, current context drifts to items at progressively earlier list positions, enabling these to be retrieved. The end result is that recency and primacy items are the most likely to be recalled, and will be recalled earliest. Mid-list items are the least likely to be recalled and will be recalled later.

However, Wixted and Rohrer (1993) point out that a distinction needs to be made between relative and absolute latency. An inverse relationship between serial output order (relative latency) and recall probability, does not necessarily mean that the mean recall latency of all items output (absolute latency) will decrease, as recall probability of those items increases. An example of this is provided by Bousfield, et al. (1954). Participants studied a list of sixty items one, two, three, four or five times. Naturally, the probability of recall increased as a function of the number of presentations. However, with increasing number of presentations, the rate of approach to asymptotic recall decreased. While latency data was not reported in this study, a slower rate of approach to asymptotic recall does suggest a longer mean latency to recall (Wixted & Rohrer, 1993).

Wixted and Rohrer (1993) attempted to explore what aspect of memory recall latency may map onto, using the Brown-Peterson paradigm. In this task, participants study trials of rapidly appearing stimuli from the same semantic category. After each trial, participants undergo a roughly thirty second distractor task to avoid rehearsal, followed by recall of the items that had been presented to them on that trial. On the final trial, the semantic category is changed. Typical findings from this paradigm, are that the percentage of items correctly recalled declines progressively across trials. On the final trial where semantic category is switched, there is a significant recovery in percentage of items correctly recalled (Wickens, 1973).

Current theoretical accounts of this paradigm centre around the phenomenon of build-up and release from proactive interference (PI). Over the first few trials, participants find it increasingly more difficult to distinguish between items presented on the current trial, and those on previous trials due to build-up of interference among items. Because of this, participants search progressively more items across trials. On the final trial, the switch in semantic category causes this interference to be released; as a result, the search for these items is much more focussed than on previous trials (Del Missier et al., 2018). Presuming that this temporal discrimination account of PI is correct, a progressive increase in recall latencies across trials would indicate that mean latency to recall indexes the breadth of memory search.

In Experiment 1 of Wixted and Rohrer (1993), participants completed four consecutive Brown Peterson trials, in the standard format as described. However, the category shift from trial 3 to trial 4 was subtle, for instance inland to coastal US states. In two of three conditions, participants received a cue for this category shift either before (before condition) or after (after condition) presentation of the fourth trial. Participants in the control condition received no cue. Accuracy data showed that the

best performance on trial 4 was in the before condition, followed by the after condition. As expected there was no release from PI in the control condition, as participants had not noticed the category shift.

For latency analysis, an exponential curve was fit to the data, as the exponential has been shown to be the most accurate representation of a sampling with replacement retrieval model (Bousfield & Sedgewick, 1944). In such a model, items are sampled from a search set which is assumed to contain items from the current list or trial, and some items from other trials. If the rate of sampling from the search set is assumed to be constant, then mean latency to recall will simply reflect the breadth of the search set. Therefore as PI builds over trials, mean latency to recall should lengthen. The exponential distribution used can be expressed thus, $f(t) = (1/\tau) e^{-t/\tau}$ where τ represents mean recall latency. Estimates of τ for the first three trials were 2.86, 4.95 and 6.61 respectively, indicating a broadening of search with the build-up of PI. On trial 4, before and after cues produced significant shortening of mean recall latency (τ = 4.62 and 4.72 respectively), indicating a release from PI and a smaller search set. Surprisingly, a shorter mean latency to recall was observed in the control condition too (τ = 5.35) although this was not significant.

Wixted and Rohrer (1993) were then interested in the source of this effect. Does the build-up of PI and the slowing of exponential retrieval arise at encoding, or during the retention interval? Temporal discrimination theory suggests that irrespective of the length of the retention interval, there should still be a broadening of memory search and lengthening of recall latencies across trials. Even at short retention intervals, participants still search more items on trial 2 than on trial 1. To address this issue, a second experiment was run, similar to Experiment 1. The procedure was largely the same, except that there was no category shift, only before cues were

presented on each trial, and trials were followed by a retention interval of either 3 seconds (short) or 27 seconds (long).

Accuracy data showed a predicted result of a significant reduction in the percentage of correctly recalled items across trials with a long retention interval, but not a short retention interval. Recall latencies were analysed differently in this experiment. It is a widely held view that reaction times in recall can be described in two phases. The first is a Gaussian stage, whereby the participant must establish a search set. The second is exponential retrieval from that search set, as described previously. The best characterisation of latencies in recall is a convolution of these two phases. This is known as an exponentially-modified Gaussian, or ex-Gaussian distribution. This distribution describes reaction times well in a number of contexts (Heathcote et al. 1991) and is expressed as in Equation 4.1:

$$f(t) = \frac{e^{-\frac{t-\mu}{\tau}+\sigma^2/2\tau^2}}{\tau\sqrt{2\pi}} \int_{-\infty}^{\frac{t-\mu}{\sigma}-\sigma/\tau} e^{-\frac{y^2}{dy}} \quad (4.1)$$

Here, τ represents the average time of the retrieval phase and μ and σ are the mean and standard deviation for duration of the search set establishment phase.

Results showed that estimates of τ significantly increased across trials for both retention intervals. The most interesting findings from this experiment concern the comparative performances of accuracy and latency with a short retention interval. In this condition there was no change in accuracy across the three trials; however, there was an increase in latency. This is significant as it demonstrates that at short retention intervals, PI does not affect a participant's ability to access correct items, but does

affect how many items they search. It should be noted that there were no ceiling effects with accuracy in this study, so a lack of difference in accuracy across trials was an interpretable finding.

There was also a significant effect of retention interval on estimates of τ for all three trials. This can be interpreted in terms of temporal discrimination theory. At short retention intervals it is relatively easy to discriminate between items on the current list and items on previous lists, therefore the search set is more focussed. With a longer retention interval, discrimination is much more challenging; therefore, the participant searches many more items. In addition to the effect on τ, there was a significant effect of retention interval on μ for trial 3. This suggests that participants may need more time to establish a search set at longer retention intervals; however, this conclusion should only be tentative as the effect was only significant on one trial.

Rohrer and Wixted (1994) followed this up by investigating recall latencies in standard-free recall. Their first study employed a simple list length manipulation where participants studied lists of words which were either three, six or nine items in length. After a twenty-second distractor period, participants were required to recall as many items as they could from the current list. Given that longer lists contain more targets it is logical to believe that the search set should also be larger. Therefore τ should increase with list length; however, there is no reason to believe that onset of recall indexed by μ should vary. Despite this rather obvious prediction, it should be noted that less overall recall does not automatically lead to a smaller search set. In the case of PI reported by Wixted and Rohrer (1993), build-up of PI leads to poorer overall recall (less items recalled and lower recall probability), but longer recall latencies.

As expected, Rohrer and Wixted (1994) found increased recall latencies and a

reduced probability of recall for longer list lengths. There was no significant effect on recall onset as predicted. A random sampling model of retrieval would adequately explain this, by asserting that a longer list increases the number of targets within a search set, hence longer recall latencies. Longer list lengths also increase the number of potentially retrievable items, which leads to greater overall recall. However this does not necessarily mean that all of these will be retrieved. Therefore, probability of recall will decline.

In a second experiment the authors manipulated study duration. While one would always expect longer study time to increase the probability of recall, the potential effect of study duration on latencies was not known. Results showed that while recall probability increased as a function of study duration, there was no effect of study duration on mean latency to recall. Again this can be explained easily by a random sampling with replacement model of retrieval. The increase in recall probability is due to a larger number of retrievable items once the search set has been established. However study duration does not affect mean latency to recall as this manipulation appears to have no impact on the number of targets and extra-list items included in the search set.

One question remained after the study duration manipulation. Could the lack of correlation between recall probability and latency be due to the range of study durations being too small? To answer this question, both list length and study duration were manipulated in a final experiment. The prediction made was that longer lists presented more slowly, would cause recall probability to increase due to additional study time, and recall latencies to lengthen owing to a longer list length.

In this experiment participants studied either six item lists presented at one item per second, or nine item lists presented at a rate of one item every four seconds.

As predicted, in the long list, long duration condition, mean latency to recall and probability of recall were significantly greater than in the short list, short duration condition. There was no significant difference between these two conditions in onset of recall.

It would appear that recall probability and accuracy measure the availability of items within, and accuracy of retrieval from the search set respectively whereas mean latency to recall indexes the breadth of the search set. Constrained search as assessed by EFR, examines how well the participant can selectively maximise the availability (and therefore recall probability) of targets relative to wrong-source items, in order to preferentially generate targets. Search accuracy is then determined by examining the number of targets and wrong source items generated. Monitoring accuracy is indexed by the ability to correctly identify if each generated item is a target or a source intrusion. Constrained search as measured by recall latency assesses the estimated search set size when recalling a single source relative to both sources. Constrained search is determined by a significantly smaller estimated search set size for recall of a single source than both sources; shorter recall latencies being indicative of a smaller search set size. If constrained search is highly accurate, then search set size for a single source should be half that of both sources.

One study that attempted to combine analyses of recall latencies and EFR to investigate memory accuracy control using a similar approach was conducted by Unsworth et al. (2013). They aimed to examine the role of proactive and retroactive interference (RI) in the control of memory accuracy. Participants were presented with either two lists or a single list of words. On the two list trials, participants were asked to recall either one of the two lists as instructed. On the single list trials they were required to recall just that list.

Experiment 2 utilised EFR in a very similar way to Chapter 2 of this thesis. At recall, participants were required to recall all of the items they could remember from either List 1 or List 2 as instructed, but also to report anything incorrect that came to mind. On the single list trials they were required to recall the single list with the same instructions. Results showed significantly higher proportion correctly recalled for the single list trials than recall of either List 1 or List 2. There was no significant difference in proportion correctly recalled between List 1 and List 2.

In a separate experiment recall latencies were analysed to observe the effects of PI and RI on search set size. The paradigm was very similar to the aforementioned. The only difference was that the participants received standard-free recall instructions for recall of List 1, List 2 or the single list as opposed to EFR. Unsworth et al. (2013) fit cumulative recall curves to the latency data of the form $F(t) = N(1-e^{-\lambda t})$, where N is total (asymptotic) recall and λ denotes the rate of approach to asymptotic recall. Here λ is the key parameter as it is often assumed that a faster rate of approach to asymptote generally indicates shorter recall latencies (Wixted & Rohrer, 1993).

Results showed that estimates of λ were greater for single list trials than either recall of List 1 or recall of List 2. There was no appreciable difference between Lists 1 and 2. Mean recall latency was also directly calculated. It was found that mean recall latency was significantly shorter for the single list trials than the two list trials. Again there was no significant difference in mean latency to recall between List 1 and List 2. Both of these experiments demonstrate clear effects of PI and RI on both the ability to constrain search, and search set size respectively; however, one issue remained untested. That is, when asked to recall one list in the presence of another, can participants filter out incorrect items from the search set? Alternatively, do

participants search through both lists in their entirety and attempt to control accuracy solely via monitoring?

A third experiment was conducted whereby participants again studied either one list or two lists of words. As with the previous experiments participants recalled either a single list or one of two lists. A third condition was added whereby participants recalled both lists in whichever order they wanted. If the participants search both lists, then latencies should be equivalent for recall of both lists, and recall of one of two lists. Latencies in both of these conditions should be longer than recall of a single list. If participants can exclude some items from the search set, then latencies should be shortest for recall of a single list, followed by recall of List 1 or 2, and longest for recall of both lists. This would indicate that participants can focus the search set to an extent, but cannot completely exclude the incorrect list.

Estimates of λ showed that rate of approach to asymptote was fastest for single list trials, followed by List 1 and List 2. Slowest rate of approach to asymptote was for recall of both lists. Analysis of mean recall latencies showed a significant main effect of list. Post-hoc comparisons revealed that mean latency for single lists was significantly shorter than List 1, List 2 and both lists. Mean latency for both lists was significantly longer than List 1, List 2 and single lists. There was no significant difference in mean latency between List 1 and List 2.

These findings provide strong support the notion that participants cannot fully focus the search set on the correct list under conditions of PI and RI, although some degree of isolation is possible. Contextual information used to reinstate the target context is inherently noisy. This causes uncertainty as to which items belong to which list, so participants cast a slightly broader search to capture as many targets as possible. The problem with this is that more irrelevant information will also be

included in the search set, increasing the likelihood of sampling an intrusion as well as a target. These findings are also an example of how EFR and latency analysis can be combined to study search processes, with both methodologies reaching the same conclusions.

The following studies aim to expand on the work conducted by Unsworth et al. (2013), by using more informative latency analyses which allow simultaneous assessment of search set size and time taken to establish a search set. Second, this work will be extended to examine focussing the search set in Mixed-lists. In previous chapters it was found that constrained search accuracy is generally poorer for Mixed-list source discriminations than for List membership. This could have been due to the contrasting roles of temporal context in these source manipulations as previously described. Regarding the issue of selective reporting in EFR, if this is not a significant confound to constrained search accuracy in EFR, the findings of the experiments in this chapter should mirror those of their EFR equivalents in Chapters 2 and 3.

However, before we can assume that these methods can be used in a complementary manner, it is necessary to examine the patterns of overt recall for both methods (for EFR this is responses written in the target box only). Recall instructions across methods are strikingly different. EFR is unconstrained recall, where participants are required to report all retrievals whether correct or incorrect whereas in the present chapter, instructions are simply to report only correct items. Bousfield and Rosner (1970) suggest that there should not be any effect of recall instruction on total correct (target) recall. However, in their study there was no instruction to explicitly monitor each retrieval. It is arguable that this monitoring instruction could lead to better output monitoring in EFR as employed in this thesis. Therefore, fewer source intrusions could be overtly output than for verbal delayed recall used for latency

analysis. To assess the impact of procedure on recall, numbers of overt targets, source intrusions, total recall and overt output dynamics will be compared. If there are no appreciable overt recall differences between the two methods, they can be seen as equivalent.

4.2 - Experiment 4.1

The first experiment was similar to Experiment 3 of Unsworth et al. (2013). The aim was to examine whether latency analysis replicated the results of Experiment 2.3, in that participants could successfully focus their search to one of two presented lists, separated by a numerical distractor task. If this is the case then estimates of search set size, tau (τ), should be greater for recall of Both lists than for recall of List 1 or list 2. Also, if participants can fully reinstate the target context, estimates of tau for List 1 and List 2 should be equal, as demonstrated using the EFR methodology of Experiment 2.3.

A secondary set of predictions concerns the onset of recall as indexed by mu (μ). It is possible that there may be a small delay in the onset of recall for recall of List 1 because participants need to locate a specific retrieval cue for targets. For recall of List 2 or Both lists, participants can use the current state of temporal context as the retrieval cue, which is inherent throughout the experiment and is therefore readily available. This prediction would manifest in a larger value of mu for recall of List 1 than recall of List 2 and both lists.

Another prediction regarding mu is borne out of CMR2 (Lohnas et al. 2015). This model has no mechanism for reinstating the context of a prior list, (or the present list) or for isolating a search set, and all attempts to locate a target are made using the current state of context as the retrieval cue. At the start of a recall period, the current state of context will most likely activate List 2 items first. If the requirement is to recall

List 1, it is likely that many source intrusions will have been retrieved before the first target. In this experiment there are no instructions to report source intrusions, therefore if CMR2 is correct and there is no context reinstatement mechanism, then there should be a larger delay in the onset of recall indexed by μ for recall of List 1 than List 2. There is no psychological reason to suspect that variability in the onset of recall as measured by sigma (σ) should differ in any case.

4.2.1 - *Methods*

4.2.1.1 - *Participants*

Thirty-six Psychology undergraduates (5 Male, 31 Female, Mean age = 21.89, *SD* = 6.23) participated in this study in return for compulsory course credit in order to pass a module.

4.2.1.2 - *Design*

The experiment had one within-subjects factor, Recall instruction. This was defined as precisely which source the participant would be required to recall on a given trial. Participants completed three experimental trials. Each consisted of two lists of ten words presented one at a time. Each list was followed by the same thirty-second digit-sorting task as used in previous chapters. Following the study phase participants were required to recall either List 1, List 2 or Both lists. Order of Recall instructions was counterbalanced by participant number. Allocation of participant numbers to a recall instruction order was conducted prior to the experiment using random sampling without replacement. Memory was tested three times over the experimental session (four including practice trial). Each test was implemented after the second digit-sorting task on each trial.

4.2.1.3 - *Materials*

Stimuli were the same sixty verbal equivalents of the Snodgrass and Vanderwart (1980) pictures that were used in Experiment 2.3, in order to provide a direct comparison between the two experiments. All stimuli in the experimental trials were presented in the centre of a computer screen, against a white background, in black Arial font; size 0.25 (PsychoPy experimental settings). Stimuli for the practice trial were ten further Snodgrass and Vanderwart (1980) pictures in their original pictorial form to avoid interference between the practice trial and experimental trials. These were presented against a white background in the centre of the computer screen, and were size [0.3, 0.5] (PsychoPy experimental settings). All participants received the same practice stimuli. Participants' recalls were recorded using an Olympus digital voice recorder (Dictaphone). The auditory buzzer to indicate the beginning of the recall period was presented through computer speakers. To account for individual differences in hearing ability, the same procedure for setting the volume as was used in experiments 2.1 - 3.1 was applied.

4.2.1.4 - *Procedure*

Study phase

The study phase was very similar to that of Experiment 2.3. Participants completed one practice trial with pictorial stimuli instead of words, and five items per list as opposed to ten. Participants were instructed to remember as many pictures as they could from both lists (forgetting previous trials), and which list each item was presented in. On each experimental trial participants viewed two lists of ten words. Each word remained on the screen for four seconds with a two second ISI. After the tenth word of List 1 participants were required to arrange the three digits that

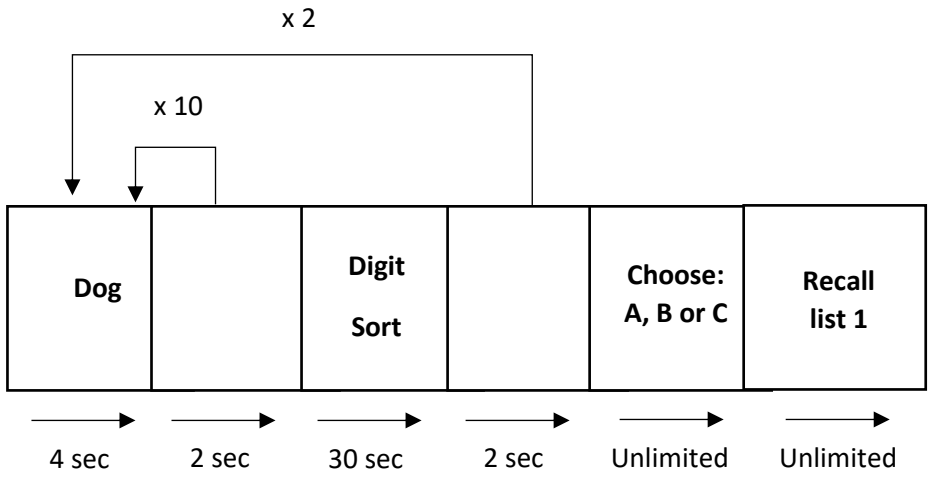
appeared on the screen into descending order. There were ten of these digit triplets over a thirty second period. This procedure was then repeated for the second list.

Test phase

The computer screen displayed a message stating “Choose: A,B or C”. A, B and C supposedly related to exactly which list/s the participant would be required to recall. In reality the order of Recall instructions for the experiment was pre-ordained by counterbalancing. A Dictaphone was also placed next to the participant at this point, and the record button pressed. After pressing the A,B or C keys on the keyboard, a message appeared on the computer screen stating one of the following: “Recall List 1”, “Recall List 2” or “Recall Both lists”, and a buzzer sounded through speakers for one second to indicate the start of the recall period. Participants verbally recalled as many items as they could from the list/s indicated by the computer screen. The experimenter ceased recording when participants verbally indicated that they could not recall any more words from the target list/s. Each individual recall period was recorded as a separate audio file (MP3). On pressing the space bar, the phrase “The next trial is about to begin. Press <space> when ready” appeared on the screen. Pressing the spacebar started the study phase of the next trial. After the third experimental trial, the words “The experiment is now over. Thank you for your participation” were displayed on the computer screen. For a schematic representation of the experimental paradigm, see Figure 4.1

Figure 4.1

Schematic Representation of the Paradigm for a single trial of Experiment 4.1.



Note. Digit sort = Digit sorting distractor task used throughout this thesis. Recall instructions could read either: Recall list 1, Recall list 2 or Recall Both lists. Participants completed all recall instructions once over the course of an experimental session. Order of recall instructions was counterbalanced as described in section 4.2.1.2.

4.2.1.5 - *Analysis*

Latency timings

Latencies were calculated using the Audacity software package. A latency was defined as the precise timing that a word was spoken during the recall period. Readings for latencies to the nearest millisecond were taken at the first sign of a fluctuation in the waveform when a word was spoken. Due to the nature of the experiment, the Dictaphone began recording prior to the beginning of the recall period. To correct for this, the lag between the beginning of the recording and the start of the recall period indicated by the auditory buzzer was subtracted from each latency.

Parameter estimation

Analysis at the individual subjects level was not possible, as recall was too poor to produce an ex-Gaussian curve. Therefore, latencies were combined from all

participants. The ex-Gaussian distribution was fitted using the R package 'retimes'. This uses a bootstrapping method to identify the parameters of a distribution. Parameter values for μ and σ are established using a Gaussian kernel estimator (Van Zandt, 2002). This is a non-parametric form of analysis whereby a Gaussian density is centred over each latency. The peak of the overall distribution (μ) is then identified as the value with the greatest density. With small sample sizes σ can often be exaggerated; therefore, the value for this parameter is calculated with respect to the mean and standard deviation of the original data, in order to ensure that only values within a theoretically plausible range are explored. This range is set using Equation 4.2:

$$\sqrt{\frac{\min(x-M)^2}{n-1}} \leq \sigma \leq S \tag{4.2}$$

Where M and S are the mean and standard deviation of the data. Once μ and σ have been Established, τ is obtained from within the calculated bootstrapped values by the method of Maximum Likelihood Estimation (MLE). The number of bootstrapped samples for each experiment will be equal to the number of participants tested.

The goal of MLE is to obtain the parameter estimates that maximise the probability of the observed data. The first step is to define the probability density function of the model of interest, in this case the ex-Gaussian function defined in Equation 1. This gives the probability of observing any data point (t) given a set of parameter values. The next step is to calculate the joint likelihood of a set of parameter values given the observed data. This is achieved by finding the product of the likelihoods for all observations in the sample. This is expressed in Equation 4.3:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{k=1}^{k} L(\boldsymbol{\theta}|y_k) \tag{4.3}$$

Where Θ is a set of parameter values, in this case values of μ, σ and τ, y is a sample of data and k represents individual data points. For ease of computation, the sum of the natural log likelihood is calculated rather than the joint likelihood. The final step is finding the best fitting parameter values. As many optimisation algorithms such as Simplex (Nelder & Mead, 1965) seek to minimise the value of the estimator, the negative sum of the log likelihood is used. This is expressed in Equation 4.4.

$$-\ln L(\boldsymbol{\theta}|\boldsymbol{y}) = -\sum_{k=1}^{K} \ln L(\boldsymbol{\theta}|\boldsymbol{y}) \tag{4.4}$$

All variables are the same as Equation 4.3. The optimisation algorithm tests various combinations of parameter values for the data in the sample. The best fitting parameter estimates are those which minimise the outcome of Equation 4.4.

To assess goodness of fit, a chi-squared analysis was conducted. Observed frequencies were obtained by partitioning the bootstrapped data into five second bins. Expected frequencies were calculated by integrating the best fitting ex-Gaussian probability density function (see equation 1). This gave the expected probabilities of observing a latency within each five second bin. Multiplying these probabilities by the total number of latencies gives the expected frequencies within each bin, given that the data precisely followed an ex-Gaussian distribution.

4.2.2 - *Results*

Overt recall data

Overt recall data from the present experiment were compared with the equivalent EFR study (Experiment 2.3) to assess whether Procedure significantly affects search and source intrusion monitoring. However, given that the present experiment and Experiment 2.3 tested different populations (General population for Experiment 2.3 and Undergraduates for the present experiment) it was necessary to check whether these populations were age matched, so that comparisons between the procedures could be effectively made. An independent *t*-test revealed that there was no significant difference in age between the two populations, $t(82) = 1.09$, $p = .28$, $d = 0.24$; however, the Bayesian evidence was just short of being conclusive, $BF_{10} = 0.39$. Therefore, it seems that these two experiments can be safely compared.

Independent *t*-tests were conducted to observe if overt target recall, overt source intrusion recall and total overt recall differed as function of Procedure. For assumed power of .8, the minimum detectable effect size was $d = 0.63$. There was no significant effect of Procedure on the number of targets overtly reported, $t(79) = 0.87$, $p = .39$, $d = 0.19$, $BF_{10} = 0.32$. In addition, there was no significant effect of Procedure on the number of source intrusions overtly reported, $t(79) = 0.45$, $p = .66$, $d = 0.10$, $BF_{10} = 0.25$. Finally there was no significant effect of Procedure on total overt recall, $t(79) = 0.78$, $p = .44$, $d = 0.18$, $BF_{10} = 0.30$. Although these analyses seem to be underpowered, Bayes Factors show credible evidence that the two procedures did not differ on any of these overt recall measures. Summary statistics for overt recall comparisons are presented in Table 4.1.

Table 4.1

Total Number of Targets and Source Intrusions Overtly Recalled and Overall Overt Recall for Verbal-Free Recall and EFR Where Source is Defined as List Membership.

Recall measure	Verbal-free recall		EFR	
	M	*SD*	*M*	*SD*
Targets	5.42	2.24	5.00	2.10
SI	0.33	0.46	0.40	0.74
Recall	5.74	2.06	5.39	1.88

Note. SI = Source Intrusions, EFR = Externalised-Free Recall, M = Mean, SD = Standard Deviation. Overt recall for EFR was calculated using items that were written in the target box only. Items that were written in the 'other' box were assumed to be withheld.

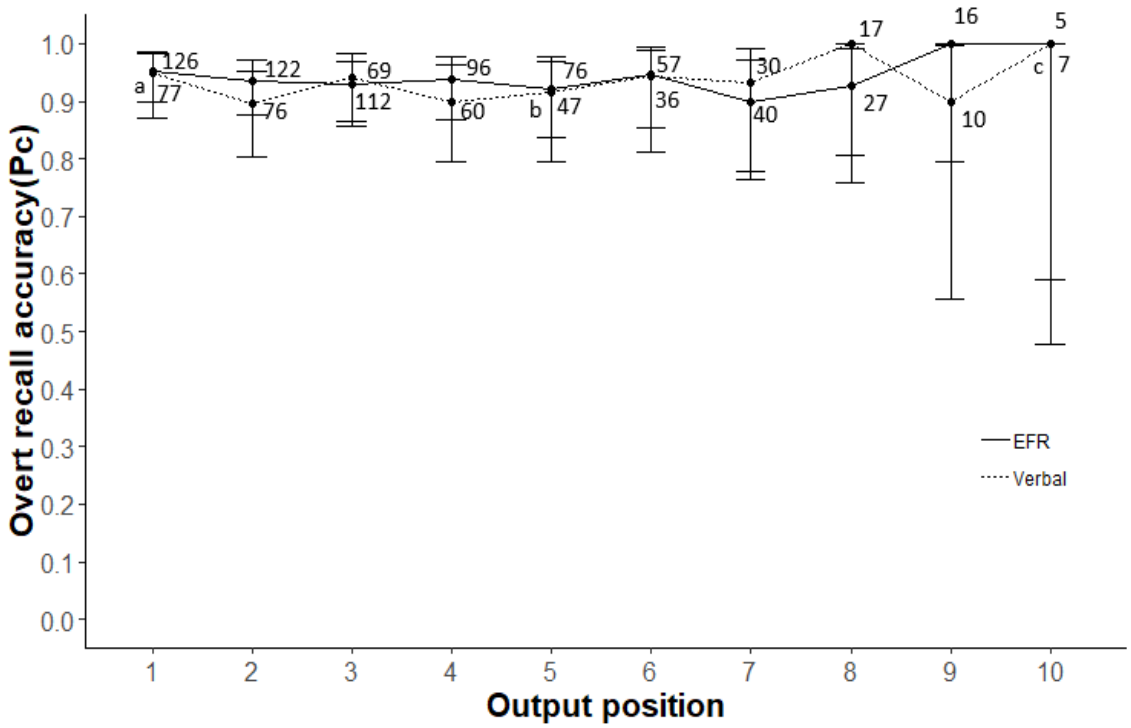
Output dynamics

Overt recall output dynamics were also compared across procedures to determine if methodology affected search processes at different stages in the recall period. As this analysis is only dealing with a subset of retrievals (target monitored) in EFR, the constraint of four items minimum recall was removed in order to maximise sample size. Because of the nature of output dynamics, sample size decreases with each output position. NHST statistics may find a null effect at late output positions simply as a result of the very small sample size. Bayes factors are far more informative for this purpose as one can quantify the evidence for an effect, or lack thereof. Therefore, only these analyses will be reported.

Bayesian contingency tables were conducted to determine if overt recall accuracy differed as a function of procedure at difference stages of the recall period. BF_{10} for early (1-3), middle (4-6) and late (7-10) output positions were 0.06, 0.09 and 0.11 respectively. This demonstrates that there is strong evidence for a lack of an effect of procedure throughout the recall period. Output dynamics are presented in Figure 4.2

Figure 4.2

Overt Recall Accuracy by Output Position for Verbal-Free Recall and EFR where Source is Defined as List Membership



Note. EFR = Externalised-Free Recall, Verbal = Verbal-Free Recall. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] EFR = 126 trials, Verbal = 77 trials

[b] Output positions 5-6: Number of EFR trials above data point. Number of Verbal trials below data point.

[c] EFR = 5 trials, Verbal = 7 trials.

A further reason for examining overt recall data is to observe whether the results of the latency analysis are in agreement with basic underlying recall processes. If latency analysis draws different conclusions to that of the behavioural data, then this brings into question the reliability of the parameter estimates.

A one-way (Recall instruction: List 1, List 2, Both lists) repeated measures ANOVA was conducted to observe if Recall instruction affected the total number of items overtly reported. Minimum detectable effect size assuming power of .8 was, η_p^2

= .04 This revealed a significant effect of Recall instruction, $F(2,70) = 43.43$, $p<.001$, $\eta_p^2 = .55$, supported by a Bayes Factor, $BF_{10} = 1.30 \times 10^{10}$. Bonferroni corrected *t*-tests (minimum detectable effect size for .8 assumed power was $d = 0.56$) revealed that there was a significant difference in total overt recall between List 1 and Both lists, $t(35) = 7.65$, $p<.001$, $d = 1.15$, and List 2 and Both lists, $t(35) = 6.66$, $p<.001$, $d = 1.01$. However, there was no significant difference in total overt recall between List 1 and List 2, $t(35) = 1.65$, $p = .10$, $d = 0.25$. Although the List 1 and List 2 comparison may be underpowered, corrected Bayesian posterior odds demonstrate evidence for a lack of an effect, in addition to support for the other comparisons, as can be seen in Table 4.3. Overall the above analyses seem to support a view that participants are able to reduce the size of their search set when they are asked to recall a single list as opposed to both lists. However, there appears to be very little evidence for a difference in breadth of search between List 1 and List 2. Summary statistics for overt recall in the different recall instruction conditions are presented in Table 4.2.

Table 4.2

Number of Targets and Source Intrusions Overtly Recalled and Total Overt Recall for each Recall Instruction in Experiment 4.1.

Recall measure	List 1		List 2		Both lists	
	M	*SD*	*M*	*SD*	*M*	*SD*
Targets	5.14	2.57	5.69	2.48	-----	-----
SI	0.31	0.67	0.36	0.54	-----	-----
Recall	5.44_b	2.37	6.03_b	2.26	9.75_{12}	4.72

Note. M = Mean, SD = Standard Deviation. Subscript letters and numbers indicate which conditions significantly differ. b = both, 1 = List 1, 2 = List 2.

Table 4.3

Bayesian Pairwise Comparisons for Total Overt Recall Across Recall Instructions in Experiment 4.1.

Level 1	Level2	Prior odds	BF_{10} uncorrected	Posterior odds
List 1	List 2	0.22	0.62	0.14
List 1	Both lists	0.22	2.16×10^6	4.85×10^5
List 2	Both lists	0.22	1.42×10^5	3.20×10^4

Note. Posterior odds are used to determine differences between groups rather than Bayes Factors.

Note that for the instruction to recall Both lists, there were no targets and source intrusions, given that there was no source discrimination required; therefore, comparisons regarding the both sources condition are only conducted on total overt recall. Paired *t*-tests were conducted to investigate whether target recall and source intrusion recall differed between Lists 1 and 2. Minimum detectable effect size with assumed power of .8 was $d = 0.48$. There was no significant difference between the two lists in the number of targets overtly reported, $t(35) = 1.43$, $p = .16$, $d = 0.22$. The effect size would suggest however that this analysis was underpowered. In addition, the Bayesian analysis was inconclusive, $BF_{10} = 0.46$. There was no significant difference between the lists in the number of source intrusions overtly reported, $t(35) = 0.42$ $p = .68$, $d = 0.05$, this time supported by a Bayes Factor, $BF_{10} = 0.19$, indicating that although underpowered it is likely that there is no difference between the lists in overt source intrusion recall. Summary statistics are presented in Table 4.2.

Latency analysis

A one-way (Recall instruction: List 1, List 2, Both lists) within-subjects ANOVA and Bayesian ANOVA were conducted on estimates of tau, to indicate if there were differences in search set size between recall of List 1, List 2 and Both lists. Minimum

detectable effect size for assumed power of .8 was η_p^2 = .04 for all repeated measures ANOVAs in this experiment. The ANOVA revealed a significant effect of Recall instruction, $F(2,70) = 1253$, $p<.001$, $\eta_p^2 = .97$, supported by a Bayes Factor, $BF_{10} = 5.63 \times 10^{72}$. Bonferroni corrected pairwise *t*-tests (minimum detectable effect size for .8 assumed power was $d = 0.56$) revealed significant differences in tau between recall of List 1 (M=9.81, SD=0.95) and recall of Both lists (M=18.30, SD=0.95), $t(35) = 41.77$, $p<.001$, $d = 8.94$, recall of List 2 (M=8.61, SD=0.74) and recall of Both lists, $t(35) = 44.03$, $p<.001$, $d = 11.39$ and recall of List 1 and recall of List 2, $t(35) = 5.77$, $p<.001$, $d = 1.42$. All pairwise comparisons were supported by the equivalent corrected Bayesian posterior odds as shown in Table 4.5.

The strongest indicator of accurate constrained search is if participants are able to reduce the size of their search set by half. Single sample *t*-tests were conducted on both List 1 and List 2 bootstrapped tau estimates, to observe if these were significantly greater than half the value of tau for both lists (9.15). Minimum detectable effect size was $d = 0.42$ for assumed power of .8. Estimates of tau were found to be significantly greater than 9.15 for List 1, $t(35) = 4.21$, $p<.001$, $d = 0.71$, supported by a Bayes Factor, $BF_{10} = 309.14$. However, tau estimates were not significantly greater than 9.15 for recall of List 2, $t(35) = -4.37$, $p = 1$, $d = 0.72$, supported by a Bayes Factor, $BF_{10} = 0.04$. This shows that although participants were able to reduce the size of their search set considerably when recalling a single list as opposed to Both lists, search appeared to be more successful when recalling List 2 than List 1.

A one-way (Recall instruction: List 1, List2, Both lists) within-subjects ANOVA and Bayesian ANOVA were conducted on estimates of mu to investigate potential differences in the onset of recall between recall instructions. The ANOVA demonstrated a significant effect of Recall instruction on mu, $F(2,70) = 19.02$, $p<.001$,

η_p^2 = .35, supported by a Bayes Factor, BF_{10} = 5.14 x 10^5. Bonferroni corrected pairwise *t*-tests (minimum detectable effect size for .8 assumed power was *d* = 0.56) revealed that the significant differences lay between List 1 (*M*=2.40, *SD*=0.43) and List 2 (*M*=2.83, *SD*=0.21), *t*(35) = 5.65, *p*<.001, *d* = 1.26 and List 1 and Both lists (*M*=2.76, *SD*=0.30), *t*(35) = 4.20, *p*<.001, *d* = 0.95. There was no significant difference in mu between recall of List 2 and Both lists, *t*(35) = 1.24, *p*=.22, *d* = 0.29. Although this final comparison could be considered underpowered, all pairwise comparisons were supported by the equivalent corrected Bayesian posterior odds as can be seen in Table 4.5. This indicates that participants initiated their search earlier when attempting to constrain search to a target list as opposed to recalling all items, but only when the target list was not the most recent list. No difference was found between onset of recall between List 2 and Both lists.

A one-way (Recall instruction: List 1, List 2, Both lists) within-subjects ANOVA was conducted to observe whether there were differences in the variability of onset of recall, indexed by sigma as a function of recall instruction. No significant differences were revealed, *F*(2,70) = 1.02, *p*=.37, η_p^2 = .03, supported by a Bayes Factor, BF_{10} = 0.22. There is a suspicion of this analysis being underpowered; however, a conclusive Bayes Factor suggests that there was no difference in sigma as a function of recall instruction. Parameter estimates for all fits from Experiment 4.1 are reported in Table 4.4.

Table 4.4

Best Fitting ex-Gaussian Parameter Estimates Across Recall Instructions in Experiment 4.1.

Parameter	List 1		List 2		Both lists	
	M	*SD*	*M*	*SD*	*M*	*SD*
τ	9.81_{2b}	0.95	8.61_{1b}	0.74	18.30_{12}	0.95
μ	2.40_{2b}	0.43	2.83_{1}	0.21	2.76_{1}	0.30
σ	0.48	0.34	0.54	0.20	0.45	0.28

Note. M = Mean, SD = Standard Deviation. Subscript letters and numbers indicate which recall instructions significantly differ. 1 = List 1, 2 = List 2, b = Both lists.

Table 4.5

Bayesian Pairwise Comparisons for τ and μ in Experiment 4.1.

Parameter	Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
τ	List 1	List 2	0.22	1.14×10^{4}	2557.96
τ	List 1	Both lists	0.22	1.02×10^{28}	2.29×10^{27}
τ	List 2	Both lists	0.22	5.88×10^{28}	1.32×10^{28}
μ	List 1	List 2	0.22	8122.57	1825.51
μ	List 1	Both lists	0.22	149.54	33.61
μ	List 2	Both lists	0.22	0.36	0.08

Note. Posterior odds are used as opposed to Bayes Factors for pairwise comparisons.

Chi-squared goodness of fit tests were conducted to assess the fit of the ex-Gaussian to the bootstrapped data. Given that each analysis had different numbers of bins and different sample sizes, the minimum detectable effect size for .8 power for List 1, List 2 and Both lists were w = 0.04, 0.05 and 0.04 respectively. These tests revealed that the best fitting ex-Gaussian differed significantly from the bootstrapped data for recall of List 1, $\chi^2(12) = 1273.80$, $p<.001$, $w = 0.15$, List 2, $\chi^2(15) = 2193.20$, $p<.001$, $w = 0.12$ and Both lists, $\chi^2(17) = 718.83$, $p<.001$, $w = 0.12$. Bayes Factors demonstrate strong evidence for departures from the ex-Gaussian for all fits. Best fitting ex-Gaussian curves are presented in Figures 4.3-4.5.

Figure 4.3

Best Fitting ex-Gaussian Curve for Recall of List 1.

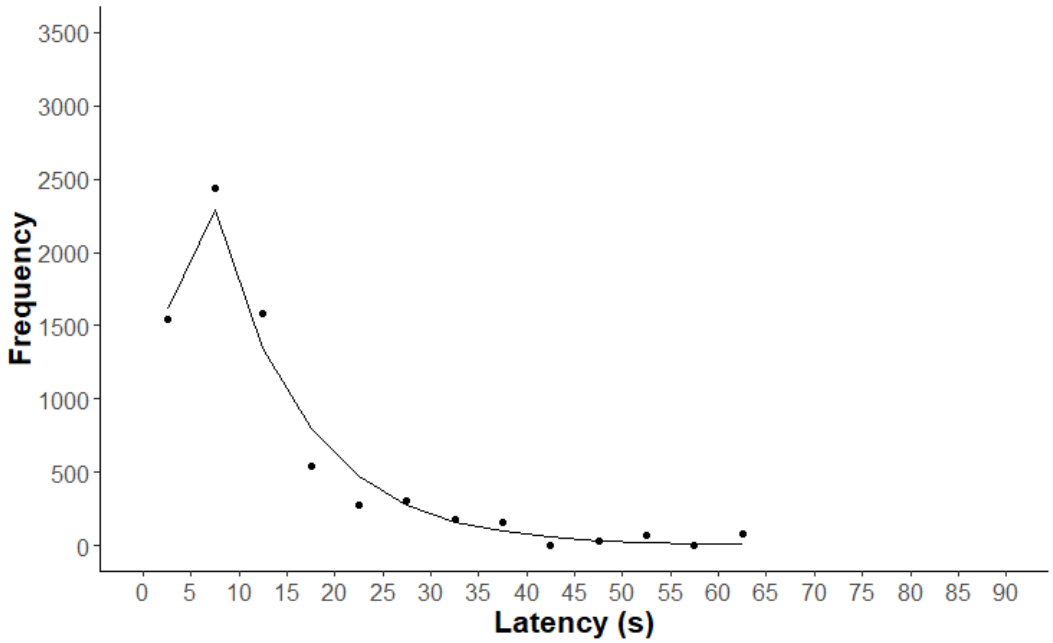


Figure 4.4

Best Fitting ex-Gaussian Curve for Recall of List 2

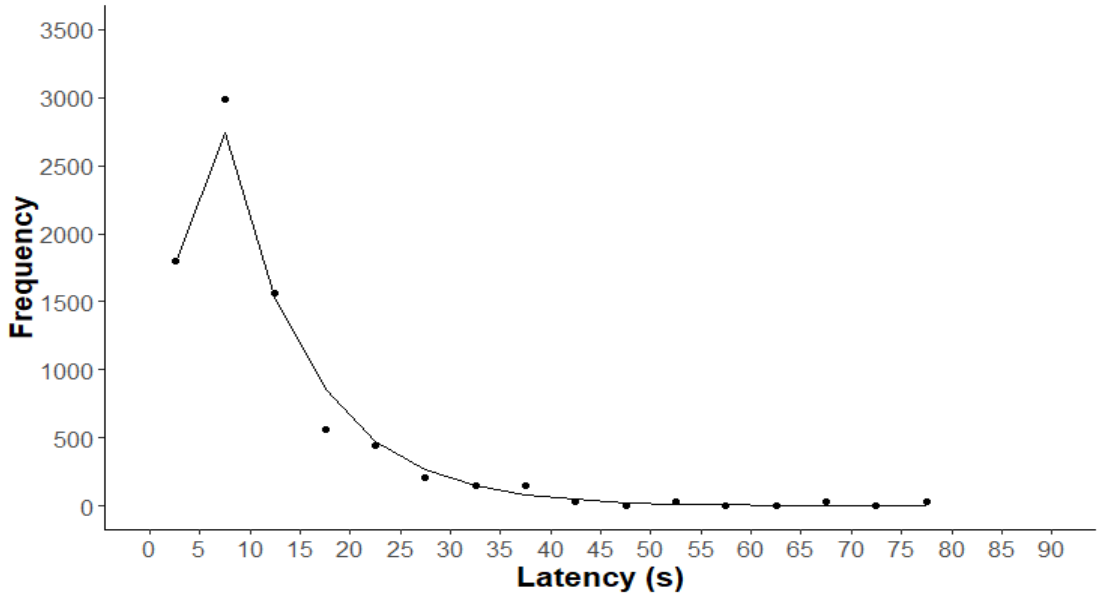
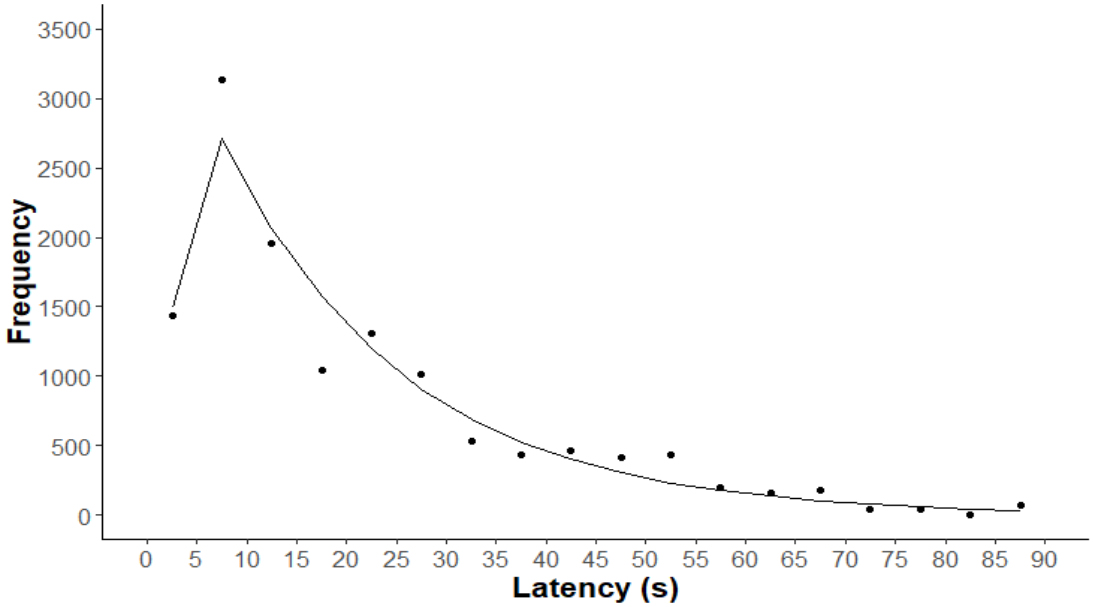


Figure 4.5

Best Fitting ex-Gaussian Curve for Recall of Both Lists.



Given the statistically poor fit to the ex-Gaussian, a chi-squared test of independence was conducted to investigate whether there is an association between recall latency and Recall instruction. This provides converging evidence for a latency difference between conditions. For all of these tests the minimum detectable effect size for .8 power was $w = 0.03$. Similar to the chi squared goodness of fit tests, latencies were partitioned into five second bins. To ensure an accurate estimate of chi-squared, bins with observed frequencies of <5 were combined with neighbouring bins, such that 40 – 60 seconds comprised two, 10 second bins, and 60 – 90 second latencies fell into a single bin.

Comparing recall of List 1 with recall of Both lists, there was a significant association between Recall instruction and recall latency, $\chi^2(10) = 1717.52$, $p<.001$, $w = 0.29$, $BF_{10} = 6.62 \times 10^{407}$. The same was true when recall of List 2 and recall of Both lists were compared, $\chi^2(10) = 2167.85$, $p<.001$, $w = 0.32$, $BF_{10} = 8.62 \times 10^{524}$. Finally

there was a significant association between Recall instruction and recall latency when List 1 and List 2 were compared, $\chi^2(10) = 110.27$, $p<.001$, $w = 0.09$, $BF_{10} = 7.53 \times 10^8$. Bayes Factors provide strong support for all of these tests.

4.2.3 - *Discussion*

The primary aims of Experiment 4.1 were to investigate if EFR and latency analyses could be usefully compared, and if so, observe if latency analysis mirrored the EFR finding of Experiment 2.3, that participants could constrain search to a single list of items when in the presence of another. Analysis of overt recall data from Experiment 2.3 and the present experiment, revealed strong evidence that procedure had no effect on number of targets, source intrusions or total number of items recalled. Furthermore, Bayesian analysis of overt output dynamics suggests that the methodologies do not differ in overt recall accuracy throughout the recall period. Taken together, these analyses indicate that comparisons across the two methodologies were viable.

Behavioural data appeared to suggest that participants could selectively search fewer items when recalling a single source than recalling both sources, as evidenced by significantly less overt recall in the Both sources condition, than for either List 1 or List 2. However, there was very little evidence that search differed between List 1 and List 2, as there was no effect of Recall instruction on source intrusions reported, overall recall, or target reporting although the latter was inconclusive. On the whole, these results support the findings of the equivalent EFR study (Experiment 2.3) which is reassuring.

Latency analysis largely reflected the findings from EFR, as estimates of tau were significantly less when recalling a single list than both lists, irrespective of

whether the requirement was to recall List 1 or List 2. This was in agreement with the behavioural data from the same experiment, suggesting that this effect was genuine.

However, a finding which did not replicate Experiment 2.3, was that constrained search was superior for List 2 than List 1. Specifically, estimated search set size was larger for recall of List 1 than List 2, demonstrating that it is more challenging to reinstate a past context than to retrieve recent items. Furthermore, participants were able to more than halve their search set size for recall of List 2, consistent with the instruction to recall half of the items; however, they were not able to do this when recalling List 1. EFR found no difference in constrained search accuracy between the two lists. It is unlikely that participants were using time of test context to recall List 1, as the difference in tau between List 1 and List 2 is much smaller than the difference between List 1 and Both lists. From this evidence it seems that to recall the most recent list participants use the time of test context as a retrieval cue which is based on recency, activating few items from previous lists. However, when retrieving a previous list, participants must reinstate the context of the target list, which produces a slightly more noisy retrieval cue that is more likely to activate non-target items.

The fact that latency methods found this effect which was absent in EFR (supported by Bayesian evidence), could indicate that latency analysis is more sensitive to constrained search than EFR, and can detect smaller more nuanced effects. However, this effect was not found in the behavioural data from verbal-free recall, as Lists 1 and 2 did not differ in target and source intrusion availability or overall recall. Therefore one must treat this finding with caution. In addition, there is ambiguity over the quality of the ex-Gaussian curve fits. Although the goodness of fit analyses indicate a large departure from the ex-Gaussian, the effect sizes for these fits are in fact quite small. Therefore, it is possible that these goodness of fit analyses are overpowered.

Despite this, the fit to List 1 is marginally poorer than the List 2 and Both list fits. Therefore, we should be particularly cautious of the estimates for List 1. Furthermore, analyses using chi-squared tests which do not rely on curve fitting appear to draw the same conclusions as the latency analyses. However, it is possible that the comparison between List 1 and List 2 may be overpowered given the highly significant P value, compared with the small effect size. Therefore, we cannot say for certain whether there was a meaningful difference between Lists 1 and 2 in breadth of search.

Secondary aims concerned the onset of recall indexed by the ex-Gaussian parameter mu. It was predicted that onset of recall may be later for recall of List 1 than List 2 or Both lists. This is because recalling List 1 requires the reinstatement of the context of the target list, and setting of appropriate retrieval cues. However, recalling List 2 and Both sources can be achieved using the inherent, evolving temporal context which exists throughout the experiment as the retrieval cue. The lack of a significant difference in mu between recall of List 2 and both lists partially supports this view. Curiously, estimates of mu were significantly larger for recall of List 2 and Both lists than List 1, which suggests that participants initiated recall earlier when a retrieval cue needed to be constructed and set, than when it is readily available, which is difficult to explain.

One cannot discount the possibility that these estimates of mu are not particularly reliable. Upon examination of Figures 4.4 and 4.5, it can be seen that the Gaussian portion of the curve, particularly the 5-10 second bin is a poor fit to the data for recall of List 2 and Both lists, which may have led to estimates of mu being somewhat misleading. Therefore, one should not attempt to draw firm conclusions based on these estimates. The next experiment will examine participants' ability to search for targets in Mixed-lists of two sources.

4.3 - Experiment 4.2

The aim of this experiment was to see if recall latency analysis supported the findings of Experiment 3.2, namely that participants can still successfully search for targets in Mixed-lists of different sources. If this is the case then estimates of tau should be significantly smaller when participants were required to recall one source than both sources. However, the difference in tau between recalling a single source and recalling both sources should be smaller than for List membership due to the unhelpful role of temporal context in Mixed-lists as discussed in Chapter 3. This should also be evident in the behavioural data, with less items being recalled for a single source than both sources. However, the difference in recall between the recall instructions should be smaller than that for List membership.

Again, the prediction that mu should be smaller for recall of two sources than one, owing to additional time taken to locate appropriate retrieval cues was tested. Sigma was not predicted to differ significantly as a function of recall instruction. There was no psychological reason to suspect significant differences in any parameter estimates between items presented at the top and items presented at the bottom of the screen. However, it was first necessary to assess whether procedure (verbal recall vs EFR) significantly affected the underlying recall data, in order to verify whether comparisons between these two methodologies are viable.

4.3.1 - *Methods*

4.3.1.1 - *Participants*

Forty further University of Plymouth Psychology undergraduates (6 Male, 34 Female, Mean age = 20.18, *SD* = 3.29) were recruited for the study, in return for compulsory course credit required to pass a module.

4.3.1.2 - *Design*

There was one between-subjects factor, Context, and one within-subjects factor, Recall instruction. Context was defined as either List membership (Experiment 4.1) or Mixed-list (screen location) discrimination. Recall instruction was defined as the items the participants were required to recall, either those presented at the top of the screen, those presented at the bottom of the screen or all of the items in the list. Stimuli were identical to those used in Experiment 4.1. On each trial participants studied a list of twenty words, half of which were presented at the top of the screen, the other half at the bottom in a random order, followed by the same digit-sorting task as previously described. Participants were then asked to recall either the items presented at the top of the screen, the items presented at the bottom of the screen or all of the items. Participants completed three trials over the course of the session, receiving each Recall instruction once. Order of recall instructions was counterbalanced by participant number. Allocation of participant numbers to a recall order was conducted prior to the experiment by means of random sampling without replacement. Memory was tested three times over the course of the experimental session (four including practice trial). Each memory test occurred after the digit-sorting task on each trial.

4.3.1.3 - *Materials*

Stimuli were identical to those used in Experiment 4.1 to provide a direct comparison. Words were printed in black Arial font; size 0.25 (PsychoPy experimental settings), and against a white background. Words appeared at either the top [0,0.5] or bottom [0,-0.5] (PsychoPy experimental settings) of the computer screen. An equal number of words were presented at the top and bottom of the screen. Participants' recalls were recorded using an Olympus digital voice recorder (Dictaphone). The

auditory buzzer to indicate the beginning of the recall period was presented in the same manner as the previous experiment, in addition to the procedure for volume setting.

4.3.1.4 - *Procedure*

Study phase

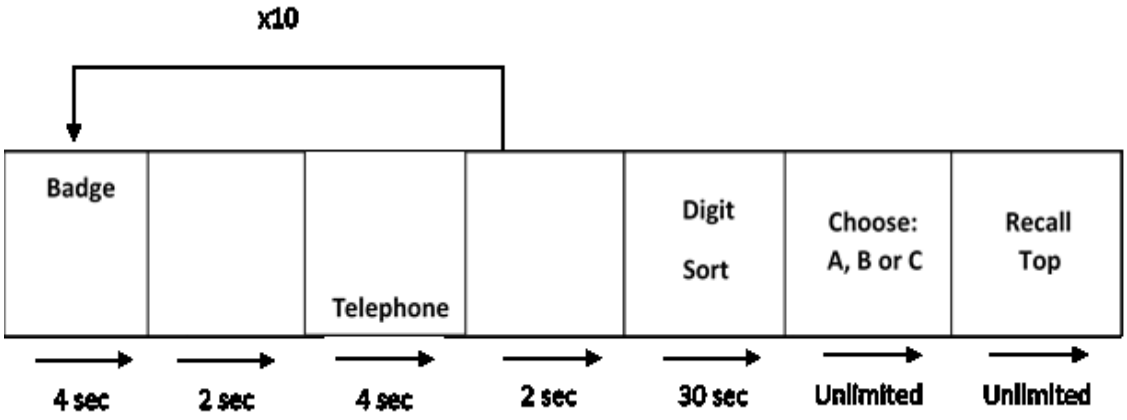
Participants were informed that for each trial they should try and remember as many items from the current trial as they could, in addition to the screen location of each word (top or bottom). The study phase had a largely identical presentation schedule as Experiment 4.1. The main difference was that the digit-sorting task only appeared after the twentieth item. The inter-stimulus interval between items ten and eleven was two seconds, the same as all other items. Ten items appeared at the top of the screen, ten items at the bottom of the screen in a random order.

Test phase

Again the test phase was largely identical to the test phase of Experiment 4.1 with the exception of the recall instructions. As opposed to being asked to recall List 1, List 2 or Both lists, participants received the instruction to either “Recall the items presented at the top of the screen”, “Recall the items presented at the bottom of the screen” or “Recall all the items”. Participants’ recalls were recorded in the same manner as Experiment 4.1, with each individual recall period as a separate audio file (MP3). See Figure 4.6 for a schematic representation of the experimental paradigm.

Figure 4.6

Schematic Representation of the Paradigm for a Single Trial of Experiment 4.2.



Note. Digit sort = Digit sorting distractor task used throughout this thesis. Recall instructions could be to recall items presented at the top of the screen, the bottom of the screen, or all the items. Participants completed all three of these recall instructions over the course of an experimental session. Allocation of screen location to items was randomised with the caveat that an equal number of items within a trial appeared at the top and bottom of the screen.

4.3.1.5 - *Analysis*

Latency timings and parameter estimation was accomplished in an identical manner to Experiment 4.1.

4.3.2 - *Results*

Overt recall data

Patterns of overt recall were examined for the present experiment and the EFR equivalent (Experiment 3.2), to observe if search and monitoring processes differed as a function of procedure. Before comparisons between experiments could be made, it was important to observe whether there were age differences between participants in the two experiments which could complicate interpretation of results. An independent *t*-test revealed that participants in Experiment 3.2 were significantly older than participants in Experiment 4.2, $t(78) = 2.97$, $p = .003$, $d = 0.66$, supported by a Bayes Factor, $BF_{10} = 9.57$. Although on first glance a medium effect size and credible evidence

from Bayesian analysis suggests that age may be a confounding factor, in reality the difference between the mean ages of participants in the two experiments was only 2.13 years. Therefore, it seems highly unlikely that age should complicate interpretation of findings.

As different numbers of items were used for the two experiments (twenty for Experiment 4.2 and twenty-four for Experiment 3.2), the data are expressed as proportions. For instance, target recall (PTarget) is the proportion of all possible targets that were overtly reported. For EFR (Experiment 3.2) only retrievals monitored as targets were included, as it is assumed that items monitored as source intrusions are withheld. The formula for PTarget for EFR is presented in Equation 4.5.

$$PTarget = \frac{t_t}{T} \tag{4.5}$$

Where t_t is the number of targets generated that were monitored as a target and T is the total number of targets in the trial. A similar formula is used to calculate the proportion of Source intrusions overtly recalled in EFR (PSI). This is expressed in Equation 4.6:

$$PSI = \frac{s_t}{S} \tag{4.6}$$

Where s_t is the number of source intrusions generated that were monitored as a target, and S is the total number of wrong-source items in the trial. Finally, the formula to calculate the proportion of the total number of items overtly recalled in EFR is expressed in Equation 4.7:

$$PRecall = \frac{t_t + s_t}{N} \tag{4.7}$$

Where t_t and s_t are the same as Equations 4.5 and 4.6, and N is the total number of items in the trial.

Independent *t*-tests were conducted to observe if there were any differences in overt target, source intrusion and overall recall between the procedures. Minimum detectable effect size for power of .8 was $d = 0.68$. There was no significant effect of Procedure on proportion of targets overtly reported, $t(67) = 0.41$, $p = .68$, $d = 0.10$, supported by a Bayes Factor, $BF_{10} = 0.27$. However, there was a significant effect of Procedure on the proportion of source intrusions overtly reported, $t(67) = 2.39$, $p = .02$, $d = 0.58$ although the Bayesian evidence was inconclusive, $BF_{10} = 2.69$. Despite this significant effect, there was no significant difference between the procedures in the proportion of all items reported irrespective of response type, $t(67) = 1.84$, $p = .07$, $d = 0.45$. The Bayesian *t*-test however, suggested that there was insufficient evidence to confirm this null effect, $BF_{10} = 1.04$.

There are suspicions of an underpowered analysis for all of these measures. Despite this, a small effect size and the Bayes Factor do suggest that target recall did not differ between the procedures. The effect size for source intrusion recall was just short of the minimum detectable effect size; however, it was still statistically significant, indicating that this effect is still observable, although not four out of five times when the alternative hypothesis is true. It is also possible that there may be a true difference in total overt recall between the procedures which could have been missed. The Bayesian analysis supports this notion as the evidence is almost completely inconclusive.

Overall, these analyses imply that there is very little evidence that procedure affected target availability, which is reassuring. However, there was an effect on reporting of source intrusions. This will be explored further in the discussion, as this

does not necessarily indicate a difference in source intrusion availability, rather differences in the quality of source monitoring between the methodologies. Overt recall summary statistics for procedural comparisons are presented in Table 4.6.

The following analyses will compare overt recall performance and recall latencies between Experiments 4.1 and 4.2. Despite the fact that these experiments drew participants from the same population (Undergraduates), it was still necessary to compare the ages of these two samples to ensure that age differences were not confounding the comparisons. An independent *t*-test revealed that there was no significant difference between the ages of the two samples, $t(74) = 1.52$, $p = .13$, $d = 0.35$; however, the Bayes Factor was inconclusive, $BF_{10} = 0.64$. Given the relatively small effect size and lack of a significant age difference between the samples, it would seem that the two experiments were comparable.

Table 4.6

Proportions of Targets, Source Intrusions and all Items Overtly Recalled For Screen Location Context Across Experimental Procedures.

Recall measure	Verbal-free recall		EFR	
	M	*SD*	*M*	*SD*
PTarget	0.46	0.19	0.44	0.25
PSI	0.11*	0.11	0.05*	0.08
PRecall	0.28	0.07	0.24	0.11

Note. PTarget = Proportion of targets recalled, PSI = Proportion of Source intrusions recalled, PRecall = Proportion of total items recalled, EFR = Externalised-Free Recall, M = Mean, SD = Standard Deviation.

* $p < .05$

A 2 (Context: Screen location, List membership) x 2 (Recall instruction: Recall single source, Recall both sources) mixed-ANOVA was conducted to investigate the effects of Context and Recall instruction on total overt recall. Assuming .8 power minimum detectable effect sizes for the main effects of Context, Recall instruction and

the interaction were η_p^2 = .07, η_p^2 =.03 and .03 respectively. There was a significant main effect of Recall instruction, $F(1,74) = 78.88$, $p<.001$, η_p^2 = .52, supported by a Bayes Factor, $BF_{10} = 1.57 \times 10^{10}$. When asked to recall both sources, participants overtly recalled more items than when the instruction was to recall a single source. However, there was no significant main effect of Context, $F(1,74) = 1.36$, $p = .25$, η_p^2 = .02, $BF_{10} = 0.39$, and no significant interaction, $F(1,74) = 3.54$, $p = .06$, $\eta_p^2 = .05$, BF_{10} = 0.97. However, it is possible that the main effect of Context and the interaction were underpowered; therefore, it would be presumptuous to conclude that these effects were absent, especially given the inconclusive Bayes Factors.

Independent *t*-tests were conducted to investigate whether target and source intrusion recall differed as a function of Context (Mixed-lists vs List membership). Both Single sources conditions were collapsed across both sources within the condition. Minimum detectable effect size for .8 power was $d = 0.63$. There was no significant effect of Context on target recall, $t(79) = 1.76$, $p = .08$, $d = 0.40$ although the analysis appears to be underpowered and the Bayesian evidence for this was inconclusive, BF_{10} = 0.89. However, participants did overtly report a significantly greater number of source intrusions in Mixed-lists than List membership, $t(79) = 3.69$, $p<.001$, $d = 0.85$, supported by a Bayes Factor, $BF_{10} = 62.41$. The implication is that when participants are required to constrain search to a single source, they include a greater number of wrong-source items in their search for Mixed-lists than List membership. There is a suggestion that this is not accompanied by fewer targets being searched for Mixed-lists, but there is insufficient evidence to draw firm conclusions about this.

Taken together, the recall data appear to show that participants can search fewer items when asked to recall a Single source than Both sources, for both Mixed-lists and List membership, although there was very little evidence that total recall

differed as a function of Context when recalling a Single or Both sources. However, there does appear to be a difference between the contexts in terms of the type of items included in the search set. When asked to recall a Single source, participants appeared to search more wrong source items for Mixed-lists than for List membership. Behavioural data for effects of context and recall instruction are presented in Table 4.7.

Table 4.7

Number of Targets and Source Intrusions Overtly Recalled and Total Overt Recall Across Contexts and Recall Instructions.

Recall	Single source Mixed		Both Sources Mixed		Single Source LM		Both Sources LM	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Targets	4.59	1.87	-----	-----	5.42	2.24	-----	-----
SI	1.08***	1.12	-----	-----	0.33***	0.46	-----	-----
Total	5.66	1.35	8.28	4.05	5.75	2.06	9.75	4.72

Note. SI = Source intrusions, LM = List membership, Total = Total overt recall, M = Mean, SD = Standard Deviation

*** $p < .001$.

Latency analysis

A paired *t*-test was conducted to observe whether search set size, as indexed by tau differed as a function of Recall instruction. Minimum detectable effect size for .8 power was $d = 0.40$. Tau was found to be significantly smaller for recall of a Single source than recall of Both sources, $t(39) = 15.24$, $p<.001$, $d = 3.61$, strongly supported by a Bayes Factor, $BF_{10} = 1.03 \times 10^{15}$. This suggests that participants were able to constrain search to a single source in Mixed-lists. A further *t*-test found that there was no significant difference in mu as a function of Recall instruction, $t(39) = 0.55$, $p=.59$, $d = 0.11$, supported by a Bayes Factor, $BF_{10} = 0.20$. A final *t*-test revealed no significant effect of Recall instruction on sigma, $t(39) = 1.20$, $p=.24$, $d = 0.26$, $BF_{10} = 0.33$. Although

the traditional analysis may have been underpowered for sigma, the Bayes factors provide conclusive evidence that Recall instruction had no effect on either the onset of recall or variability in the onset of recall.

A 2 (Context: Screen location, List membership) x 2 (Recall instruction: Single source, Both sources) mixed-ANOVA was conducted to observe whether Context affected participants' ability to control search set size. In this case, the between-subjects factor was Context and the within-subjects factor was Recall instruction. The ANOVA revealed a significant interaction between Context and Recall instruction, $F(1,78) = 949.72$, $p<.001$, $\eta_p^2 = .92$, supported by a Bayes Factor, $BF_{10} = 1.82 \times 10^{66}$. The minimum detectable effect size with assumed power of .8 for this interaction was $\eta_p^2 = .03$. Bonferroni corrected simple main effects analyses (minimum detectable effect size with assumed power of .8 for all simple main effects analyses in this experiment was $d = 0.70$) were conducted to observe if Experiment/Context affected tau at either level of Recall instruction. When recalling a Single source, estimates of tau were significantly larger for Mixed-lists ($M = 10.32$, $SD = 0.60$) than List membership ($M = 9.22$, $SD = 0.59$), $t(78) = 8.58$, $p<.001$, $d = 1.92$. However, estimates of tau were significantly smaller for Mixed-lists ($M = 12.65$, $SD = 0.68$) than List membership ($M = 18.30$, $SD = 0.95$) when recalling Both sources, $t(78) = 16.16$, $p<.001$, $d = 3.61$. This suggests that Context has a significant bearing on search set size, which will be explored further in the discussion. These were supported by corrected Bayesian posterior odds for the equivalent analyses, (see Table 4.9).

A further 2 (Context: Screen location, List membership) x 2 (Recall instruction: Single source, Both sources) mixed-ANOVA was conducted to determine if Context affected differences in the onset of recall between Recall instructions. This revealed a significant interaction between Context and Recall instruction, $F(1,78) = 4.89$, $p = .03$,

η_p^2 = .06, supported by a Bayes Factor, BF_{10} = 3.22. Minimum detectable effect size for this interaction with assumed power of .8 was η_p^2 = .03. Bonferroni corrected simple effects analyses revealed that there was no significant difference in mu between List membership (*M* = 2.61, *SD* = 0.25), and Mixed-list context (*M* = 2.54, *SD* = 0.16) when recalling a Single source, *t*(78) = 1.66, *p* = .10, *d* = 0.37, although this may be underpowered. Onset of recall was significantly later for List membership (*M* = 2.76, *SD* = 0.30) than Mixed-lists (*M* = 2.52, *SD* = 0.17) when recalling Both sources, *t*(78) = 4.47, *p*<.001, *d* = 1.00. These were supported by the equivalent Bayesian corrected posterior odds (see Table 4.9). Therefore, onset of recall was earlier for Mixed-lists only when both sources were being recalled.

A final 2 (Context: Screen location, List membership) x 2 (Recall instruction: Single source, Both sources) mixed-ANOVA was conducted to observe if Context had any significant impact on the differences in variability of recall onset time across recall instructions, indexed by sigma. The ANOVA revealed a significant interaction between Context and Recall instruction, *F*(1,78) = 5.58, *p*=.02, η_p^2 = .07, supported by a Bayes Factor, BF_{10} = 4.37. Minimum detectable effect size for this interaction with assumed power of .8 was η_p^2 = .03. Bonferroni corrected simple main effects analyses revealed that variability in onset of recall was greater for List membership (*M* = 0.56, *SD* = 0.21) than Mixed-list context (*M* = 0.45, *SD* = 0.15) when participants were required to recall a Single source, *t*(78) = 2.87, *p*=.005, *d* = 0.64. However, this effect size is lower than the minimum detectable effect size indicating that this effect may not be observable four times out of five when the null hypothesis is true. There was no significant difference in sigma between List membership (*M* = 0.45, *SD* = 0.28) and Mixed-list context (*M* = 0.49, *SD* = 0.15) when Both sources were recalled, *t*(78) = 0.75, *p*=.46, *d* = 0.17. Again, these were supported by the equivalent analyses of corrected Bayesian

posterior odds (See Table 4.9). Parameter estimates for mu, sigma and tau, and comparisons with Experiment 4.1 are presented in Table 4.8.

Table 4.8

Best Fitting ex-Gaussian Parameter Estimates for Both Recall Instructions in Mixed-List and List Membership Contexts.

Param	Single source Mixed		Both sources Mixed		Single source LM		Both sources LM	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
τ	**10.32_{sl}**	0.60	**12.65_{bl}**	0.68	9.22_{sm}	0.59	18.30_{bm}	0.95
μ	2.54	0.16	2.52_{bl}	0.17	2.61	0.25	2.76_{bm}	0.30
σ	0.45_{sl}	0.15	0.49	0.15	0.56_{sm}	0.21	0.45	0.28

Note. Bold text denotes where significant differences lie between recall instructions in Experiment 4.2. These same analyses for the List membership are not reported here as a more fine grained analysis of this experiment is presented in Table 4.3. Subscript text indicates significant simple main effects of Context. sl =Single List membership, bl = Both List membership, sm = Single Mixed, bm = Both Mixed, Param = Parameter, LM = List membership, M = Mean, SD = Standard Deviation.

Table 4.9

Bayesian Simple Main Effects Analyses for Comparisons Between Experiments 4.1 (List Membership) and 4.2 (Mixed-Lists).

Param	Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
τ	Single source Mixed	Single source LM	0.41	7.81×10^{9}	3.24×10^{9}
τ	Both sources Mixed	Both sources LM	0.41	1.54×10^{23}	6.39×10^{22}
μ	Single source Mixed	Single source LM	0.41	0.76	0.32
μ	Both sources Mixed	Both sources LM	0.41	720.44	298.42
σ	Single source Mixed	Single source LM	0.41	7.51	3.11
σ	Both sources Mixed	Both sources LM	0.41	0.30	0.12

Note. Param = Parameter, LM = List membership.

Chi-squared goodness of fit tests were conducted to assess whether the ex-Gaussian curves were a good fit to the bootstrapped data. Data were combined for items presented at the top and bottom of the screen, as they were not expected to differ for any parameter. Minimum detectable effect size for .8 power for the Single source and Both source fits were $w = 0.03$ and $w = 0.04$ respectively. This analysis demonstrated that the best fitting ex-Gaussian differed significantly from the bootstrapped data for recall of a Single source, $\chi^2(15) = 1136.30$, $p<.001$, $w = 0.11$. This was also found for recall of Both sources, $\chi^2(15) = 614.61$, $p<.001$, $w = 0.11$. Best fitting ex-Gaussian curves are presented in Figures 4.7 and 4.8. Despite the poor statistical fit, the effect sizes for the goodness of fit tests can be considered small; therefore, these analyses may be overpowered, and the fits may not be as poor as they initially seem. Figures 4.7 and 4.8 also demonstrate that the data largely follow the general pattern of the best fitting ex-Gaussian; therefore, the ex-Gaussian may be appropriate.

Figure 4.7

Best Fitting ex-Gaussian Curve for Recall of a Single Source in Mixed-Lists.

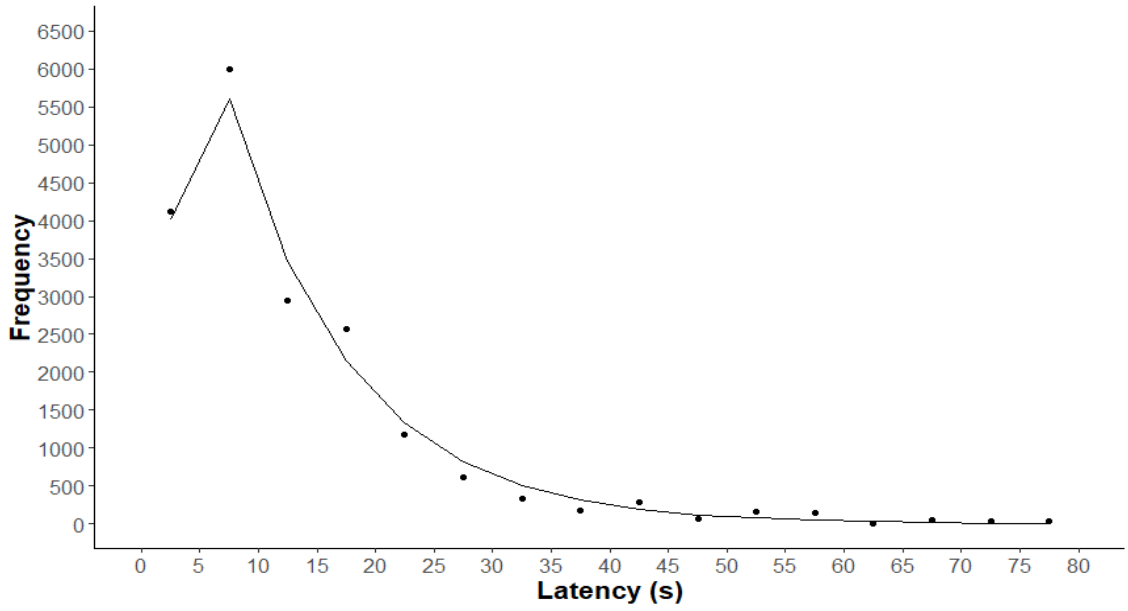
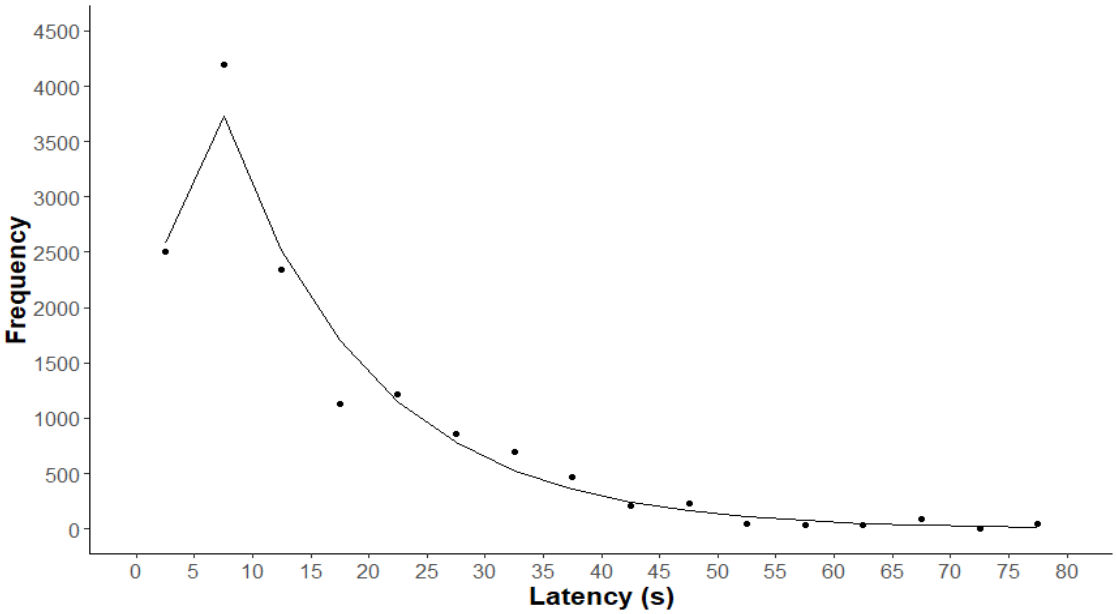


Figure 4.8

Best Fitting ex-Gaussian for Recall of Both Sources in Mixed-Lists.



Again, given that the chi-squared goodness of fit test revealed a poor fit to the ex-Gaussian distribution, the distributions for recalling half of the items and recalling all of the items were compared with a chi-squared test of independence. Data were again partitioned into 5 second bins. To ensure that no bin contained a frequency of <5, the bins encompassing 60 – 80 seconds were expanded to cover 10 second intervals. Minimum detectable effect size for .8 power was $w = 0.02$. The chi-squared test revealed that there was a significant relationship between latency and Recall instruction, $\chi^2(13) = 1324.07$, $p<.001$, $w = 0.20$, supported by a Bayes Factor, $BF_{10} = 7.27 \times 10^{269}$.

4.3.3 - *Discussion*

The purpose of this experiment was to investigate whether participants could constrain search by source features (spatial location) in Mixed-lists. To ensure comparisons could be usefully made between EFR and verbal-free recall, overt recall across the methodologies was investigated. Reassuringly there was little evidence that

target availability or overall item availability was affected by procedure. However, participants overtly recalled significantly more source intrusions in verbal-free recall than EFR. In fact, this may not be such an issue, as this does not necessarily mean that incorrect source items are more accessible in verbal-free recall. It could be the case that the explicit monitoring instruction in EFR leads to better monitoring of wrong source items; hence, fewer source intrusions being reported. It would be far more concerning if target availability was affected by procedure as target monitoring is near ceiling in EFR studies conducted in previous chapters; therefore, any differences in target reporting would almost certainly be due to target availability. Fortunately, this was not the case. Source intrusion monitoring is much less accurate; therefore, there is certainly scope for this difference in source intrusion reporting across procedures to be a monitoring effect. As the present exercise only deals with search processes, this increase in source intrusion reporting may ultimately be unimportant.

It was predicted that participants could constrain search in Mixed-lists; however, constrained search would be poorer for Mixed-lists than for List membership due to the contrasting effects of temporal context on incorrect item activation as discussed in Chapter 3. Latency analysis appears to support this. When parameter estimates were examined, search set size as indexed by tau was significantly larger for recall of both sources than recall of a single source. In addition, chi-squared analysis revealed an association between latency and recall instruction. This was supported by the behavioural data, which shows that participants overtly recalled fewer items when recalling a single source than both sources. Furthermore, in line with predictions, search set size was larger for Mixed-lists than List membership when participants recalled a single source, indicating a broader search for Mixed-lists than List membership. The behavioural data appear to support this, as participants overtly

recalled more source intrusions in Mixed-lists than List membership when recalling a single source. The difference between tau for recall of a single source and both sources was also smaller for Mixed-list contexts than List membership. Although this was not directly tested in the behavioural data, the difference in total recall between recall of one source and both sources was numerically larger for List membership than for Mixed-lists. Given the similar sample size for the two experiments, this likely represents a larger effect for List membership.

One interesting finding was that search set size was significantly larger for List membership than Mixed-lists when recalling both sources. Moreover, this effect was far larger than the difference between contexts when recalling a single source, demonstrating that the majority of the difference between List membership and Mixed-list contexts can be attributed to fewer items being searched when all items in the trial are to be recalled. This is also reflected in the behavioural data, as the difference in recall between the two contexts was numerically although not statistically, larger in the both sources than the single source condition.

This would appear to point to screen location being a weaker retrieval cue than temporal context. A weaker retrieval cue leads to a less targeted search. This is evidenced by the smaller difference in tau between the Single source and Both source condition for Mixed-lists than List membership, and a larger search set size for Mixed-lists in the Single source condition. This can also be seen in the behavioural data as less targets and more source intrusions were recalled for Mixed-lists than List membership. A weaker retrieval cue would also lead to a smaller search set size in the Both sources condition for Mixed-lists than List membership, which is what was observed. Additional evidence for this comes from greater recall (albeit not statistically significant) for List membership than Mixed-lists in the Both sources condition. An

explanation based on temporal context being unhelpful in retrieving targets for Mixed-lists is unlikely, as it would not account for the differences in tau and recall between the two contexts in the Both sources condition.

One further hypothesis was made, in that mu should be less for recall of both sources than a single source, due to additional time needed to set an appropriate retrieval cue. No evidence for this was found. However, participants did initiate search later for List membership than for Mixed-lists when they were required to recall both sources. It is not clear why this should be the case, as theoretically participants should not have to set an appropriate source cue when both sources are to be recalled. Search may be initiated using the current state of experimental context, (Polyn et al. 2009a) as source accuracy is not a consideration.

Finally, it was found that variability in the onset of recall was greater for List membership than for Mixed-lists when a single source was recalled. Again, it is not immediately obvious why this should be the case. A potential explanation could be related to the manner in which the data were analysed. The single source data in Experiment 4.2 were the combination of items presented at the top and bottom of the screen. There is no reason to suspect that there would be any difference in onset of recall between these two sources. However, for Experiment 4.1 participants were required to recall items presented in two different time periods. Indeed Table 4.2 illustrates that participants initiated recall significantly later for recall of List 2 than recall of List 1. The increased variability in the onset of recall for List membership may be due to the combination of data from List 1 and List 2, for the comparison with Mixed-lists. On the whole, the latency analysis and behavioural data from Experiment 4.2 suggest that EFR does not appear to suffer from excessive selective reporting

confounds. The final experiment of this chapter will explore a factor which may influence the success of constrained search in Mixed-lists.

4.4 - Experiment 4.3

Once it had been established that latency measures replicate the EFR findings that participants could constrain search in Mixed-lists, factors which influence the ability to constrain search could then be explored. This experiment will be a latency measures replication of Experiment 3.1. To recap, this experiment investigated the effects of Source Similarity on constrained search and monitoring. Participants studied four trials of words. In two of these trials the words were presented either through headphones or on the screen in a random order (Low-similarity), or in a male or female voice in a random order (High similarity). EFR findings demonstrated that as expected, participants were worse at monitoring the source of items in the High-similarity trials than the Low-similarity trials. However, slightly surprisingly, this finding did not extend to constrained search. Participants could selectively retrieve targets in both the High and Low-similarity trials, but interestingly, Similarity had no effect on the ability to constrain search, both in terms of raw number of targets and source intrusions retrieved, and recall dynamics.

The present experiment aimed to replicate the findings of Experiment 3.1. It was expected that estimates of tau would be significantly less when recalling one source than two sources in both Similarity conditions. However, Similarity should have no effect on the difference in tau between recall of one source and recall of both sources. There was no expectation that mu or sigma would differ as a function of Similarity. As neither Experiment 4.1 or 4.2 supported the hypothesis that mu would be greater for recall of a single source than both sources, there was no expectation that this experiment would yield anything different.

4.4.1 - *Methods*

4.4.1.1 - *Participants*

Forty further participants were recruited for this experiment in return for compulsory course credit required to pass a module. Two participants were removed from the analysis for failure to recall a single correct item in one condition. A further two were removed for recalling a greater number of source intrusions than targets in at least one condition, indicating either an inability to distinguish between target and wrong source items or failure to follow instructions. Therefore, the sample size was thirty-six (4 Male, 32 Female, Mean age = 20.18, *SD* = 3.29).

4.4.1.2 - *Design*

There were two within-subjects factors, Source Similarity and Recall instruction. The former was defined as the stimulus presentation format. Eighty words were randomly allocated to one of four experimental trials. In the High-similarity trials, words were presented either in a male voice or a female voice in a random order. In the Low-similarity trials, words were presented either through speakers or on the screen in a random order. Participants studied two consecutive trials for each Similarity condition. The order of these conditions was counterbalanced so that half of the participants received the High-similarity trials first, and the other half received the Low-similarity trials first. The Recall instruction factor was defined as the source or sources that the participant would be required to recall (Single source or Both sources). This was also counterbalanced by participant number so that the two Recall instruction orders (Single, Both and Both, Single) would appear an equal number of times in each Similarity condition. Participant numbers were allocated a Similarity condition order and a Recall instruction order prior to the experiment by means of

random sampling without replacement. In addition the thirty-second digit-sorting-distractor task was reinstated after the tenth item of each trial to aid comparison with Experiment 3.1. Memory was tested four times over the course of the experimental session (five with practice trial). Recall periods were implemented after the second digit sorting task on each trial.

4.4.1.3 - *Materials*

This experiment used identical stimuli to those in Experiment 3.1, so that the two experiments could be effectively compared. Visual stimuli in the Low-similarity condition were presented on a computer screen in black Arial font; size 0.25 (PsychoPy experimental settings), against a white background. For all auditory stimuli in both Similarity conditions the computer screen was blank throughout stimulus presentation. Auditory stimuli for both Similarity conditions were the same audio files used for Experiment 3.1. The voice used for auditory stimuli in the Low-similarity condition was the same as the male voice in the High similarity condition. Stimuli were allocated to trials such that no individual word appeared in more than one trial for any given participant. For instance the word 'horse' could not appear as an auditory stimulus in the Low-similarity condition, and again as female spoken word in the High-similarity condition for the same participant.

To account for individual differences in hearing ability, volume was adjusted manually to suit the participant prior to the experiment, in the same fashion as Experiment 3.1. by presenting a series of beeps of different volumes through headphones. Due to covid restrictions the experimenter could not do this for the participants. Therefore, they were instructed how to do this. Stimuli for the practice trials were ten Snodgrass and Vanderwart (1980) images in their original pictorial form. These were presented in the centre of the screen, with a height and width of 50% full

screen size. All participants received the same practice stimuli. Due to social distancing regulations, collection of participants' recalls using a Dictaphone was not possible. Instead this was achieved by participants speaking their recalls into a microphone connected to a computer at the other end of the room, so the experimenter did not need to approach the participant. The experimenter started recall recordings after the second digit-sorting task, and terminated recordings when the participant stated that they could not remember any more words. All four recall periods for each participant were recorded as a single audio file (MP3).

4.4.1.4 - *Procedure*

Study phase -

Before the first trial of the Low-similarity condition, participants were told that for each trial, they should remember as many words as they could from both lists (forgetting words from previous trials), in addition to how each word was presented (through the headphones or on the computer screen). They were then presented with five words through the headphones (auditory), and five words in the centre of the computer screen (visual) in a random order one at a time. Each word had a presentation duration of four seconds, with a two second inter-stimulus interval (ISI). As with the equivalent EFR study (Experiment 3.1) all auditory stimuli irrespective of Similarity condition were of subtly different duration; therefore, stimuli were presented over a four second period, with silence filling time when words were not being spoken. Following the tenth word, participants completed the digit-sorting task as described earlier (black text; Arial font; size 0.25 PsychoPy experimental settings; white background). Then, the final ten auditory and visual words (five auditory, five visual) were presented randomly. The digit-sorting-distractor task then appeared again for thirty seconds. The High-similarity condition was almost identical. Study

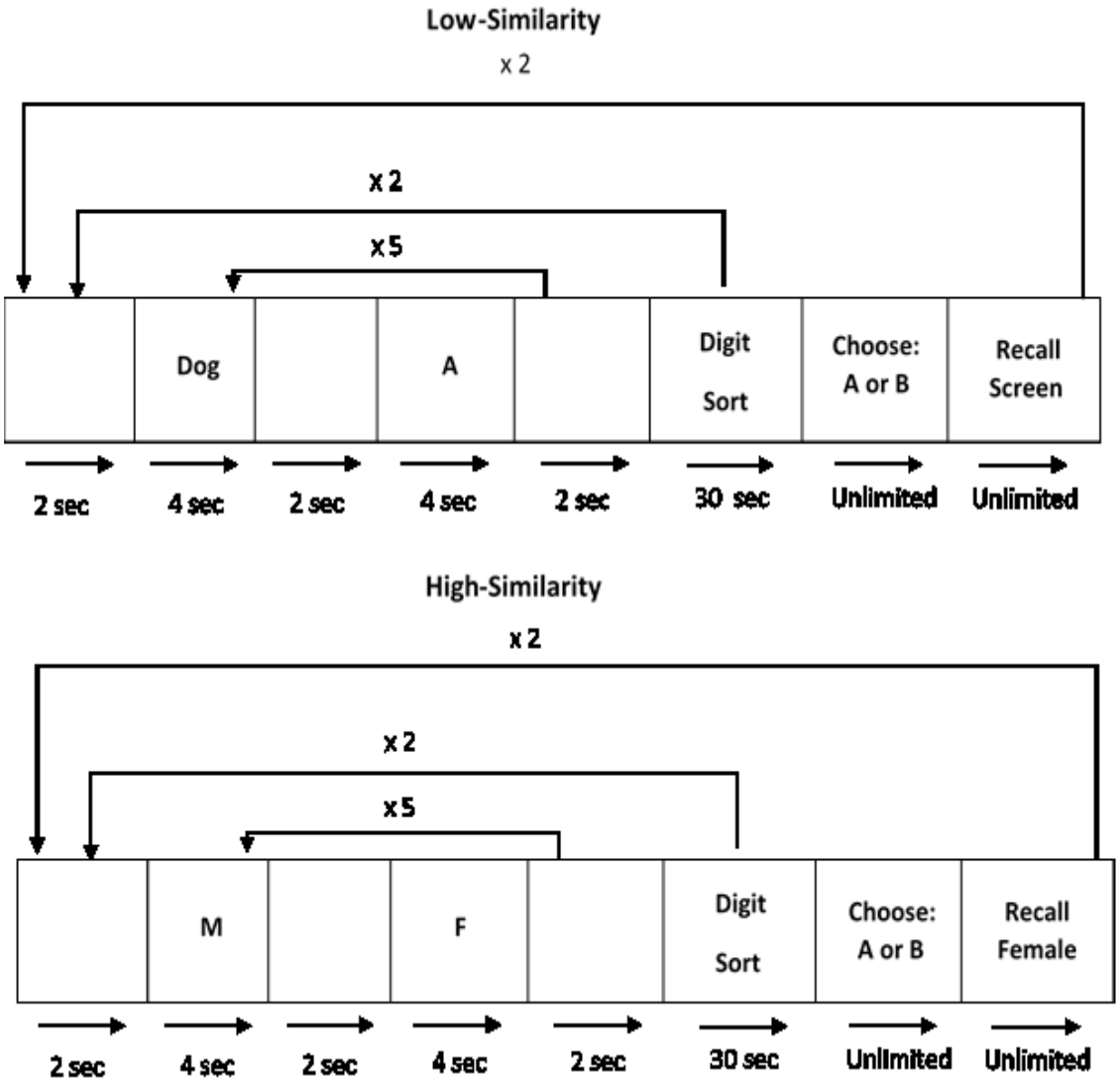
instructions were that participants should remember as many words as they could from both lists (forgetting all previous trials), and to remember how each word was presented (in a male voice or a female voice). The only difference in stimulus presentation was that all stimuli were presented through the headphones, half of the words in a male voice and the other half in a female voice.

Test phase -

The test phase was largely identical to Experiments 4.1 and 4.2. One difference was the recall instructions. In the High-similarity trials, the participants received the instruction to “Recall the items spoken in a male voice”, “Recall the items spoken in a female voice” or “Recall all the items”. In the Low-similarity trials, instructions were to “Recall the items you saw on the screen”, “Recall the items you heard through the speakers”, or “Recall all the items”. For the single source trials (recall male/female, recall auditory/visual) the source participants needed to recall was randomised. Upon hearing the auditory beep accompanying the instruction, participants spoke their responses into the microphone. See Figure 4.9 for a schematic representation of the experimental paradigm.

Figure 4.9

Schematic Representation of the Paradigm for Experiment 4.3.



Note. Digit sort = Digit sorting distractor task used throughout this thesis. A = Auditory word, Low-similarity condition, M = Male spoken word, High-similarity condition, F = Female spoken word, High-similarity condition. Recall instructions for the Low-similarity condition could be to Recall items presented through the headphones (single source), Recall items presented on the screen (single source), or to Recall all the items (both sources). Recall instructions for the High-similarity condition could be to Recall items spoken in a Male voice (single source), Recall items spoken in a Female voice (single source) or recall all items (both sources). As there were two trials per condition, participants only had one single source recall instruction and one both sources recall instruction.

4.4.1.5 - *Analysis*

Latency timings and parameter estimation was accomplished in an identical fashion to the previous two experiments.

4.4.2 - *Results*

Overt recall data

As with Experiments 4.1 and 4.2, overt recall data were compared with the equivalent EFR experiment (Experiment 3.1) to investigate potential differences in search and monitoring processes across the procedures. In this experiment, the data were further subdivided into Low and High-similarity conditions to observe if procedure differentially affected the two Similarity conditions. Before these comparisons could be made, it was important to check for potential age differences between the 2 samples which might complicate interpretation of the results. An independent t-test revealed that there was no significant age difference between the samples, $t(98) = 0.49$, $p = .62$, $d = 0.10$, supported by a Bayes Factor, $BF_{10} = 0.24$. The very small effect size, lack of a significant age effect and complementary Bayesian evidence suggests that these two experiments are comparable.

Three 2 (Procedure: Verbal-Free Recall, EFR) x 2 (Similarity: High, Low) mixed-ANOVAs were conducted to observe whether overt target recall, overt source intrusion recall and total overt recall respectively were affected by experimental procedure and Similarity. Assuming .8 power, minimum detectable effect size for the main effects of Procedure and Similarity were $\eta_p^2 = .06$ and $\eta_p^2 = .02$ respectively. Minimum detectable effect size for the interaction was $\eta_p^2 = .02$. The first ANOVA demonstrated that there was no significant main effect of Procedure on target recall, $F(1,91) = 1.63$, $p = .21$, $\eta_p^2 = .02$; however, a Bayes Factor suggests that there is not sufficient evidence to support the null, $BF_{10} = 0.49$, and there are suspicions of low power. Additionally,

there was no significant main effect of Similarity on target recall, $F(1,91) = 1.59$, $p = .21$, $\eta_p^2 = .02$ although there was not quite sufficient evidence to confirm this, $BF_{10} = 0.33$. Finally there was no significant interaction between Experiment and Similarity, $F(1,91) = 1.22$, $p = .27$, $\eta_p^2 = .01$ although again the Bayesian evidence was not quite conclusive, $BF_{10} = 0.38$. From this it seems unlikely that Procedure or Similarity affected target recall; however, we cannot be certain, as the main effect of Procedure appears to be underpowered. Bayes factors were also inconclusive despite indicating that there is more evidence for a lack of an effect than an effect in all cases.

A second 2 (Procedure: Verbal-Free Recall, EFR) x 2 (Similarity: High, Low) mixed-ANOVA revealed that there was a significant main effect of Procedure on the number of source intrusions overtly reported, $F(1,91) = 13.72$, $p<.001$, $\eta_p^2 = .13$, supported by a Bayes Factor, $BF_{10} = 33.98$. Bonferroni corrected *t*-tests (minimum detectable effect size for .8 assumed power was $d = 0.62$) found that more source intrusions were overtly reported during verbal-free recall than EFR in the High-similarity condition, $t(91) = 2.68$, $p = .009$, $d = 0.57$. This was also true of the Low-similarity condition, $t(91) = 2.66$, $p = .009$, $d = 0.57$. Although effect sizes for both of these comparisons were less than the minimum detectable effect size, meaning that these effects may not be found four times out of five when the alternative hypothesis is true. However, corrected Bayesian posterior odds support both of these comparisons as can be seen in Table 4.11. There was no significant main effect of Similarity on the number of source intrusions reported, $F(1,91) = 2.58$, $p = .11$, $\eta_p^2 = .03$, although the Bayes Factor suggests that there is only anecdotal evidence for a lack of an effect, $BF_{10} = 0.61$. There was no significant interaction between Procedure and Similarity, $F(1,91) = 0.01$, $p = 1.00$, $\eta_p^2 < .001$, supported by a Bayes Factor, $BF_{10} = 0.20$. From this analysis we can see that participants overtly reported more source intrusions

during verbal-free recall than EFR. However, while participants did overtly report more source intrusions in High-similarity conditions for both procedures, this was not enough to be significant, although we cannot be certain of this due to the inconclusive Bayes Factor. This again suggests poorer source monitoring in verbal-free recall.

A third 2 (Procedure: Verbal-Free Recall, EFR) x 2 (Similarity: High, Low) mixed-ANOVA was conducted to assess the effect of Procedure and Similarity on the total number of items overtly recalled. There was found to be no significant main effect of Procedure, $F(1,91) = 0.003$, $p = .96$, $\eta_p^2 < .001$, supported by a Bayes Factor, $BF_{10} = 0.25$, no significant main effect of Similarity, $F(1,91) = 0.21$, $p = .65$, $\eta_p^2 = .002$, supported by a Bayes Factor, $BF_{10} = 0.17$, and no significant interaction between Procedure and Similarity, $F(1,91) = 1.27$, $p = .26$, $\eta_p^2 = .01$; however, the Bayes Factor was inconclusive, $BF_{10} = 0.36$. Therefore, we can say that neither Procedure nor Similarity had any appreciable effect on overall item availability. Bayes Factors demonstrate that there is good evidence for making this assertion. Summary statistics for overt recall data are presented in Table 4.10.

Table 4.10

Number of Targets and Source Intrusions Overtly Recalled and Total Overt Recall in Each Similarity Condition Across Procedures.

Recall measure	High-similarity EFR		High-similarity Verbal		Low-similarity EFR		Low-similarity Verbal	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Targets	4.60	1.82	3.91	2.12	4.69	1.97	4.49	1.84
SI	0.35_{hv}	0.59	0.77_{he}	0.94	0.17_{lv}	0.35	0.60_{le}	1.14
Recall	4.95	1.68	4.69	2.14	4.86	1.92	5.09	1.92

Note. SI = Source intrusions, EFR = Externalised-Free Recall, M = Mean, SD = Standard Deviation. Subscript letters indicate where significant differences lie. hv = High-similarity Verbal, he = High-similarity EFR, lv = Low- similarity Verbal and le = Low-similarity EFR.

Table 4.11

Bayesian Pairwise Comparisons for Main Effect of Procedure on Overt Source Intrusion Recall in Both Similarity Conditions.

Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
High-similarity EFR	High-similarity Verbal	0.41	9.81	4.06
Low-similarity EFR	Low-similarity Verbal	0.41	8.06	3.34

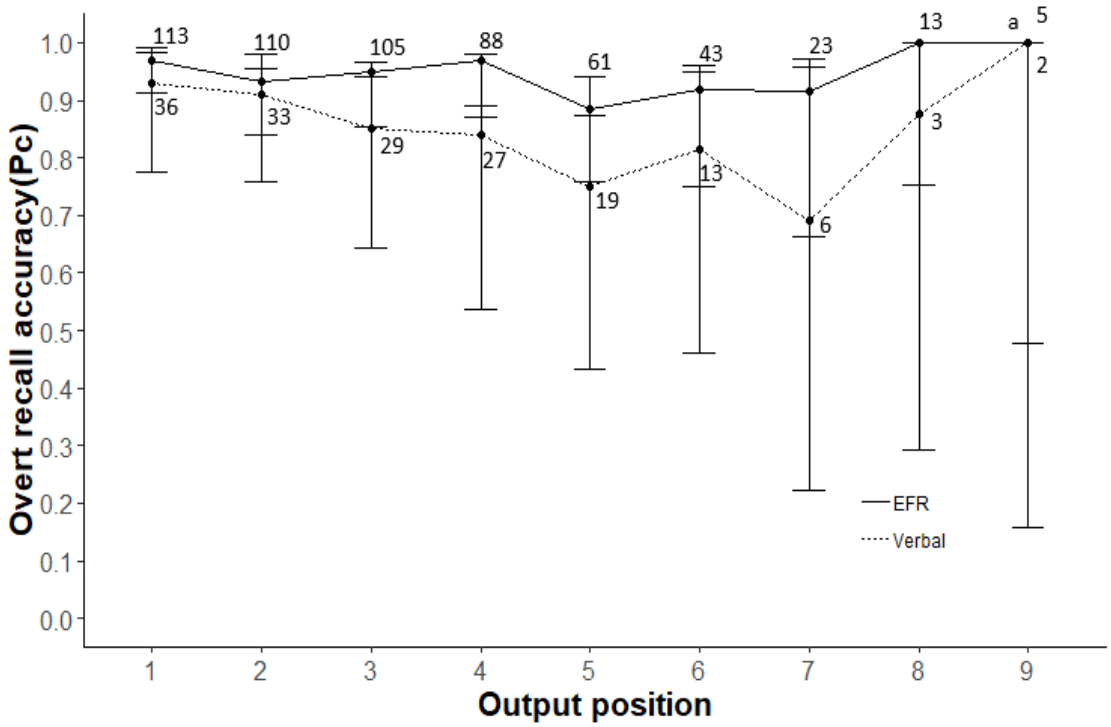
Note. EFR = Externalised-Free Recall, Verbal = Verbal-Free Recall. Posterior odds are used to infer effects rather than Bayes Factors.

Output dynamics

As with Experiment 4.1, overt recall output dynamics were examined to investigate potential differences in recall accuracy between the methodologies at early (1-3), middle (4-6), and late (7-9) output positions, separated into the two Similarity conditions. Again only Bayes Factors are reported for the reasons previously stated. In the High-similarity condition BF_{10} for early (1-3), middle (4-6) and late (7-9) output positions were 0.23, 29.05 and 0.47 respectively. Therefore, there was evidence for no difference between the procedures early in the recall period, evidence for a difference mid-way through the recall period, and weak evidence for no difference at late output positions. As can be seen from Figure 4.10, the trend appears to be an increasing difference in recall accuracy between the procedures as the recall period progresses.

Figure 4.10

Overt Recall Accuracy by Output Position for Both Procedures in the High-Similarity Condition.



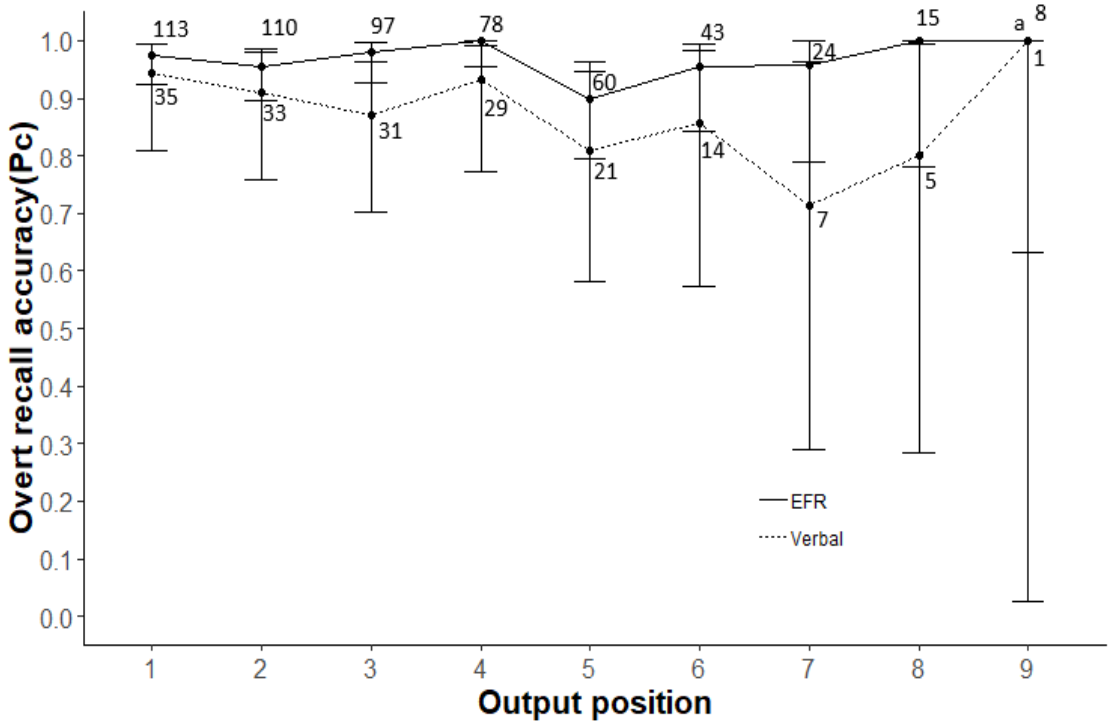
Note. EFR = Externalised-Free Recall, Verbal = Verbal-Free Recall. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] EFR = 5 trials, Verbal = 2 trials.

For the Low-similarity condition, BF_{10} values for early (1-3), middle (4-6) and late (7-9) output positions were 1.01 respectively, 2.81 and 3.29 respectively. This indicates no credible evidence for either hypothesis at early output positions, weak evidence for a difference between the procedures mid-way through the recall period, and evidence for a difference at late output positions. Figure 4.11 shows a similar pattern of recall accuracy to the High-similarity condition. However the performance difference between the two procedures appears by visual inspection to be slightly smaller.

Figure 4.11

Overt Recall Accuracy by Output Position for Both Procedures in the Low-Similarity Condition.



Note. EFR = Externalised-Free Recall, Verbal = Verbal-Free Recall. Error bars represent 95% confidence intervals at each output position. Digits above/below each data point indicate the number of trials contributing data to that output position.

[a] EFR = 8 trials, Verbal = 1 trial.

A 2 (Similarity: High, Low) x2 (Recall instruction: Single source, Both sources) within-subjects factorial ANOVA was conducted to examine the effects of Similarity and Recall instruction on total number of items overtly reported. Assuming .8 power, minimum detectable effect size for these main effects and the interaction was $\eta_p^2 = .04$ There was a significant main effect of Recall instruction, with participants recalling a greater number of items when recalling both sources compared to a single source, $F(1,34) = 50.84$, $p<.001$, $\eta_p^2 = .60$, $BF_{10} = 1.68 \times 10^{10}$. There was no significant main effect of Similarity, $F(1,34) = 1.52$, $p = .23$, $\eta_p^2 = .04$, $BF_{10} = 0.26$. There was also no significant interaction between the factors, $F(1,34) = 0.04$, $p = .85$, $\eta_p^2 = .001$, $BF_{10} =$

0.24. From traditional and Bayesian analyses, it would seem that participants were recalling fewer items when asked to recall a single source than both sources; however, there is strong evidence that recall was not affected by Similarity.

Paired *t*-tests were also conducted to investigate whether target and source intrusion recall in the Single source conditions were affected by Similarity. Minimum detectable effect size was $d = 0.48$. It was found that there was no significant difference in target recall as a function of Similarity, $t(34) = 1.48$, $p = .15$, $d = 0.29$, however the Bayesian analysis was inconclusive, $BF_{10} = 0.49$. There was also no significant effect of Similarity on source intrusion recall, $t(34) = 0.71$, $p = .48$, $d = 0.16$, supported by a Bayes Factor, $BF_{10} = 0.23$. Therefore, it seems that we can only confidently say that Similarity had no effect on source intrusion recall. We should not attempt to draw conclusions about target recall due to low power and an inconclusive Bayes Factor.

As with Experiment 3.1, the male voice used for the auditory source in the Low similarity condition was the same as the male source in the High-similarity condition. Therefore, it was important to investigate potential interference across conditions. This was explored by examining the participants' trial 3 and trial 4 recall outputs for instances of items presented in the male voice on trials 1 and 2. For each participant, the total number of these interference intrusions across trials 3 and 4 were divided by the total number of recalled items in trials 3 and 4, to gain the proportion of these interference errors in the recall output (P_{int}). See Equation 3.1 in Appendix B. A P_{int} score significantly greater than 0 would indicate a degree of across condition interference. Mean P_{int} for the present experiment was .02, (*SD* = .04) A single sample *t*-test (minimum detectable effect size assuming .8 power was $d = 0.42$) was conducted to investigate whether P_{int} was significantly greater than 0, indicating interference. It

was found that this was the case, $t(35) = 2.14$, $p = .02$, $d = 0.36$; however, the Bayes Factor was inconclusive, $BF_{10} = 2.65$.

On the surface it seems that there may be significant across condition interference; however, the obtained effect size was lower than the minimum detectable effect size for .8 power, meaning that the effect may not be detectable four times out of five when the alternative hypothesis is true. In addition, the inconclusive Bayes Factor suggests that on the whole, there is insufficient evidence to be certain of interference. Finally, it is important to note that for verbal-free recall not all generated items are output; therefore, it is difficult to ascertain the true proportion of interfering items in the search set, as multiple items (interfering or otherwise) will be filtered out by monitoring.

When these results are combined with the total free recall data presented above it would seem that participants can constrain their memory search in both Similarity conditions; however, there is very little evidence to suggest that Similarity itself had any effect on search. Recall data from the present experiment is presented in Table 4.12.

Table 4.12

Targets and Source Intrusions Overtly Recalled and Total Overt Recall by Similarity and Recall Instruction in Experiment 4.3.

Measure	High-similarity Single source		High-similarity Both sources		Low-similarity Single source		Low-similarity Both sources	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Targets	3.91	2.12	-----	-----	4.49	1.84	-----	-----
SI	0.77	0.94	-----	-----	0.60	1.14	-----	-----
Recall	4.69	2.14	7.97	3.85	5.09	1.92	8.51	3.70

Note. There was a significant main effect of Recall instruction on Recall, such that participants recalled more items for both sources compared to a single source. M = Mean, SD = Standard Deviation

Latency analysis

A 2 (Similarity: High, Low) x2 (Recall instruction: Single source, Both sources) within-subjects factorial ANOVA was conducted to investigate if Similarity had an effect on estimated search set size, as indexed by tau. Minimum detectable effect size assuming .8 power for all effects was $\eta_p^2 = .04$. This revealed a significant interaction between Similarity and Recall instruction, $F(1,39) = 88.14$, $p<.001$, $\eta_p^2 = .69$ supported by a Bayes Factor, $BF_{10} = 3.72 \times 10^{14}$. Simple main effects were examined using Bonferroni corrected paired *t*-tests (minimum detectable effect size for .8 assumed power was $d = 0.58$), to determine where the significant differences in tau originate. In the High-similarity trials, estimates of tau were significantly smaller for recall of a Single source ($M = 10.79$, $SD = 1.32$) than Both sources ($M = 11.89$, $SD = 0.65$), $t(39) = 4.40$, $p<.001$, $d = 1.06$. However, unexpectedly in the Low-similarity trials, estimates of tau were lower when recalling Both sources ($M = 10.76$, $SD = 0.86$) than recalling a Single source ($M = 12.50$, $SD = 0.93$), $t(39) = 8.80$, $p<.001$, $d = 1.94$. This suggests that participants could successfully constrain search in High-similarity trials. However in Low-similarity trials, there were more items in the search set when recalling a single source than both sources.

When recalling a single source, tau was significantly less in High-similarity trials ($M = 10.87$, $SD = 1.32$) than in Low-similarity trials ($M = 12.50$, $SD = 0.93$), $t(39) = 7.45$, $p<.001$, $d = 1.50$. However, when recalling both sources tau was significantly greater for the High-similarity trials ($M = 11.89$, SD = 0.65) than the Low-similarity trials ($M = 10.76$, $SD = 0.86$), $t(39) = 6.31$, $p<.001$, $d = 1.48$. These suggest that when recalling a single source, estimated search set size was smaller in the High-similarity trials compared to Low-similarity trials. However, the opposite was true when recalling both

sources. All pairwise comparisons were supported by the equivalent corrected Bayesian posterior odds (see Table 4.14).

A further 2 (Similarity: High, Low) x2 (Recall instruction: Single source, Both sources) within-subjects ANOVA was conducted to investigate effects of Similarity and Recall instruction on initiation of recall, indexed by mu. There were no significant main effects of Similarity, $F(1,39) = 0.06$, $p=.81$, $\eta_p^2 = .002$ $BF_{10} = 0.17$, or Recall instruction, $F(1,39) = 1.90$, $p=.18$, $\eta_p^2 = .05$ $BF_{10} = 0.33$. There was also no significant interaction between the factors $F(1,39) = 1.37$, $p=.25$, $\eta_p^2 = .03$, $BF_{10} = 0.02$. Bayesian evidence supports all of these null results. Combined, these findings imply that there was no effect of Similarity or Recall instruction on onset of recall.

A final 2 (Similarity: High, Low) x2 (Recall instruction: Single source, Both sources) within-subjects ANOVA was conducted to investigate potential effects of Similarity and Recall instruction on variability in onset of recall, indexed by sigma. This revealed no significant main effect of Similarity, $F(1,39) = 0.75$, $p=.39$, $\eta_p^2 = .02$. Although underpowered, the Bayes Factor supports a lack of a main effect of Similarity, $BF_{10} = 0.24$. However, there was a significant main effect of Recall instruction, $F(1,39) = 16.17$, $p<.001$, $\eta_p^2 = .29$, $BF_{10} = 21.84$ with greater variability in onset of recall for recalling Both sources than a Single source. There was no significant interaction between Similarity and Recall instruction, $F(1,39) = 0.02$, $p=.90$, $\eta_p^2 <.001$, $BF_{10} = 0.23$. Bayes Factors provide strong support for these assertions. Descriptive statistics are presented in Table 4.13.

Table 4.13

Best Fitting ex-Gaussian Parameter Estimates for Both Similarity and Recall Instructions Conditions in Experiment 4.3.

Param	High-similarity Single source		High-similarity Both sources		Low-similarity Single source		Low-similarity Both sources.	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
τ	**10.87**$_{ls}$	1.32	**11.89**$_{lb}$	0.65	12.50$_{hs}$*	0.93	10.76$_{hb}$*	0.86
μ	2.33	0.15	2.42	0.27	2.36	0.22	2.36	0.32
σ	0.20	0.16	0.31	0.22	0.23	0.25	0.35	0.30

Note. Bold text indicates a simple main effect of Recall instruction in the High-similarity trials, Asterix indicates a simple main effect of Recall instruction for the Low-similarity trials. Subscript lettering denotes where simple main effects of Similarity lie, ls = Low Single, lb = Low Both, hs = High Single, hb = High Both. The main effect of Recall instruction on sigma is not shown (greater σ for recall of both sources than a single source). Param = Parameter, M = Mean, SD = Standard Deviation.

Table 4.14

Bayesian Simple Main Effects Analysis for Tau in Experiment 4.3

Parameter	Level 1	Level 2	Prior odds	BF_{10} uncorrected	Posterior odds
τ	High-similarity Single source	High-similarity Both sources	0.41	289.92	120.09
τ	Low-similarity Single source	Low-similarity Both sources	0.41	1.21×10^8	4.99×10^7
τ	High-similarity Single source	Low-similarity Single source	0.41	1.30×10^6	5.38×10^5
τ	High-similarity Both sources	Low-similarity Both sources	0.41	8.00×10^4	3.32×10^4

Note. Posterior odds rather than Bayes Factors are utilised to infer effects.

Chi-squared goodness of fit tests were conducted to evaluate how well the best

fitting ex-Gaussian fit the bootstrapped data. Minimum detectable effect sizes assuming .8 power for the High-similarity Single source, High-similarity Both sources, Low-similarity Single source and Low-similarity both sources fits were $w = 0.06$, $w = 0.04$, $w = 0.05$ and $w = 0.04$ respectively. Observed latency frequencies differed significantly from the best fitting ex-Gaussian for the High-similarity Single source fit, $\chi^2(22) = 26199$, $p<.001$, $w = 0.20$ the High-similarity Both sources fit, $\chi^2(15) = 646.34$, $p<.001$ $w = 0.11$, the Low-similarity Single source fit $\chi^2(19) = 1498.70$, $p<.001$, $w = 0.12$ and the Low-similarity Both sources fit, $\chi^2(15) = 2194.90$, $p<.001$, $w = 0.14$. Again, despite the poor mathematical fits, the data followed the basic shape of the ex-Gaussian, justifying it's adoption as the distribution of choice. Fits are presented in Figures 4.12-4.15.

Figure 4.12

Best Fitting ex-Gaussian Curve for the High-Similarity, Single Source Condition.

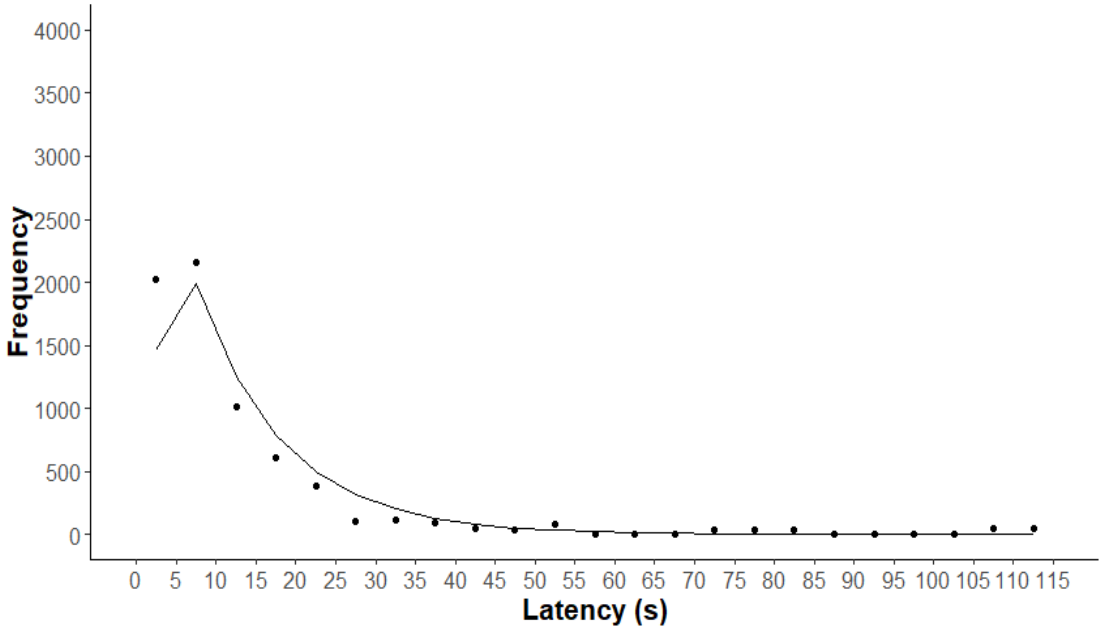


Figure 4.13

Best Fitting ex-Gaussian Curve for the High-Similarity, Both Sources Condition.

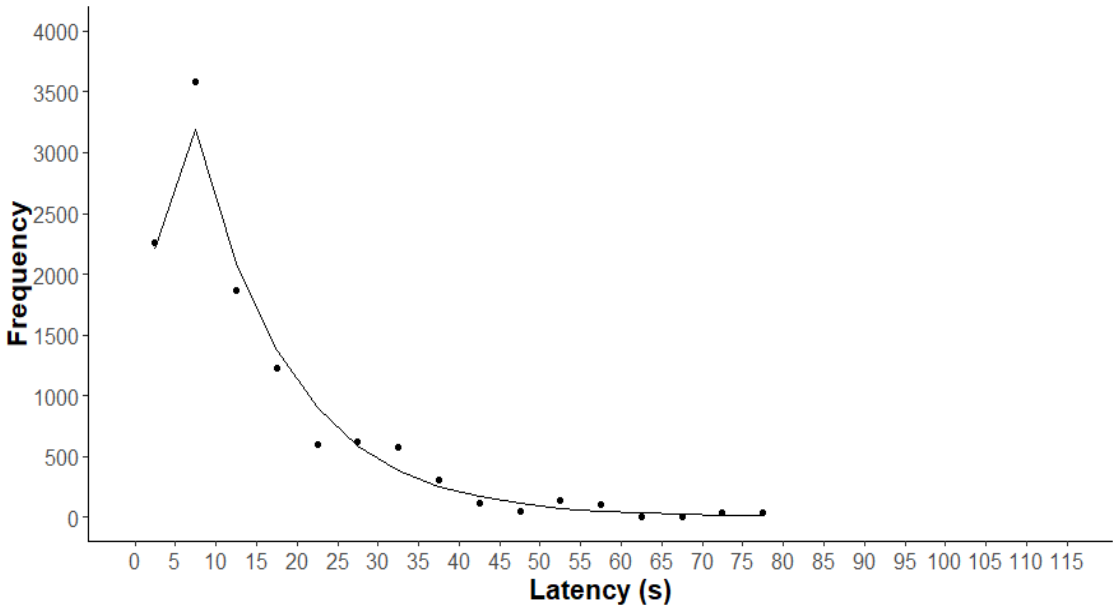


Figure 4.14

Best Fitting ex-Gaussian Curve for the Low-Similarity, Single Source Condition.

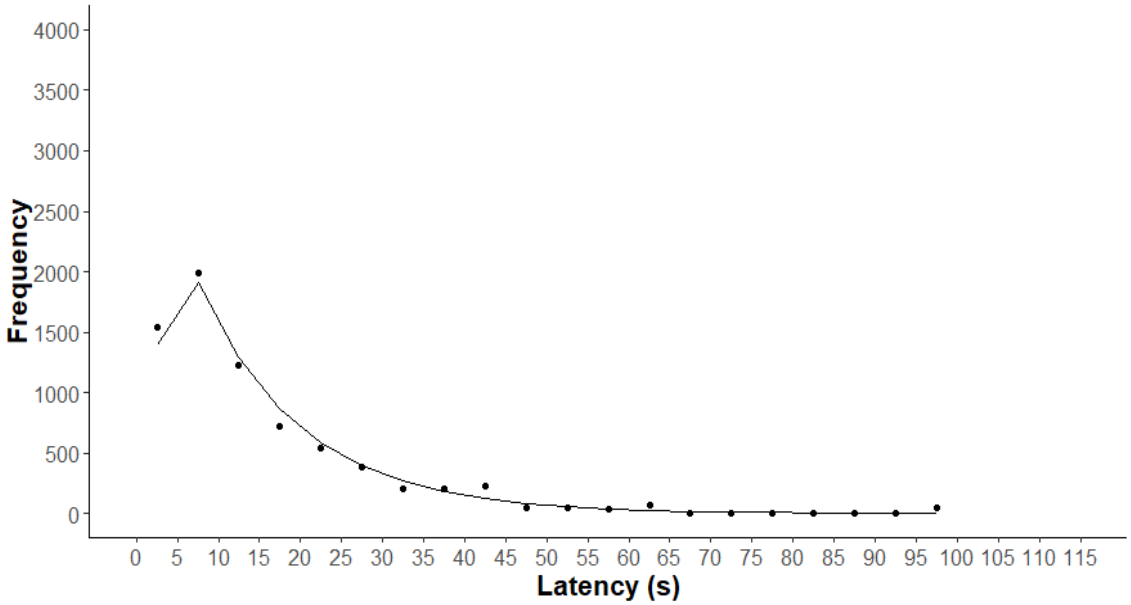
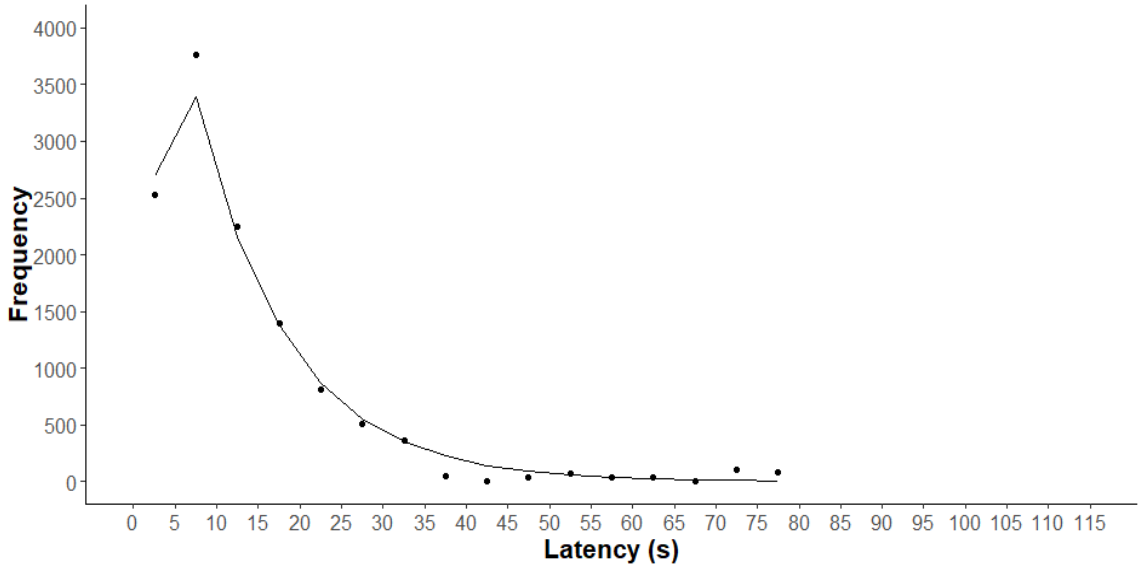


Figure 4.15

Best Fitting ex-Gaussian Curve for the Low-Similarity Both Sources Condition.



Chi-squared tests of independence were conducted to compare distributions. As with the previous two experiments, data were partitioned into 5 second bins. However to ensure all bins had a frequency >5, bins were combined between 40 and 50 seconds, 55 and 75 seconds, and 75 and 115 seconds. Minimum detectable effect sizes assuming .8 power for all chi-squared analyses were $w = 0.03$. Chi-squared tests revealed a significant relationship between Recall instruction and latency in the High-similarity trials $\chi^2(11) = 687.67$, $p<.001$, $w = 0.19$, $BF_{10} = 6.04 \times 10^{140}$, and Low-similarity trials, $\chi^2(11) = 613.36$, $p<.001$, $w = 0.18$, $BF_{10} = 6.79 \times 10^{117}$. Furthermore, when participants were asked to recall half of the items, there was a significant relationship between Similarity and latency, $\chi^2(11) = 542.18$, $p<.001$, $w = 0.20$, $BF_{10} = 1.04 \times 10^{106}$. Finally, when participants were required to recall all items, there was a significant relationship between Similarity and latency, $\chi^2(11) = 454.29$, $p<.001$, $w = 0.14$, $BF_{10} = 1.04 \times 10^{106}$. Bayes Factors provided extremely strong support for these analyses.

4.4.3 - *Discussion*

The aim of this experiment was to investigate the effects of Source Similarity on

ability to constrain search. The original hypotheses were that participants should be able to constrain search in both High and Low-similarity trials, and that constrained search should be poorer when sources are more similar. However EFR findings from Experiment 3.1 showed that Similarity had no effect on constrained search accuracy. Therefore, it was expected that if there were no issues with selective reporting for EFR, estimates of tau should not differ as a function of Similarity.

First, the overt recall data were examined to see if the two methodologies could be viably compared. It seems unlikely that experimental procedure affected search, as there was no main effect of Procedure on total overt recall, and no significant effect of Procedure on target availability, although the evidence for this was not strong. As with Experiment 4.2 there was strong evidence that participants output more source intrusions in verbal free recall than EFR which may indicate superior source monitoring in EFR. Without knowing the precise number of source intrusions generated, we cannot say for certain the extent to which this is a monitoring effect. However, this is much more likely to be diagnostic of source monitoring than target reporting, as target monitoring is near ceiling in the previous EFR studies conducted in this thesis. Therefore, the number of targets reported is likely to be almost identical to the number of targets generated. Source intrusion monitoring is considerably less accurate; therefore, source intrusions reported may differ from the number of source intrusions generated, and is likely to index at least in part, quality of monitoring. On the whole these findings are largely reassuring, as search processes appeared to be very similar across the procedures, indexed by no significant difference in targets reported, accompanied by no difference in overall recall. This is of crucial importance for the present treatment, as latency analysis only indexes search processes. Therefore, the significant difference between the methodologies in source intrusions

reported is not problematic.

The patterns of overt recall dynamics for both Similarity conditions appear to show increasingly superior recall accuracy for EFR as the recall period progresses. This is predictable given that there are very few source intrusions generated early in the recall period, so one would not expect the procedures to differ. As the recall period progresses the number of source intrusions generated increases due to a falling base rate of targets. Therefore, source intrusion monitoring becomes more important at later output positions. If we assume that search processes across the two procedures are the same, then output accuracy will be increasingly driven by source monitoring as the recall period progresses. The recall data appear to show poorer source monitoring in verbal-free recall; therefore, this explains why the overt recall accuracy performance gap between the procedures widens as source monitoring becomes more prominent.

Regarding the effects of Similarity on overt recall accuracy in the two procedures, there was good evidence that Similarity did not affect overall item availability, indicating that Similarity appears to have no effect on search for either procedure. This is promising, as it demonstrates that the EFR finding of no effect of Similarity on search is replicable in a more orthodox constrained free-recall task. There was also no main effect of Similarity on the number of source intrusions output; however, the Bayes Factor suggested that there was insufficient evidence to draw firm conclusions about this. This does not necessarily indicate that Similarity had no effect at all on source monitoring, especially given that Experiment 3.1 found superior source intrusion monitoring for Low-similarity than High similarity lists. Rather, it may be the case that there were simply not enough source intrusions overtly reported for the difference to be detectable. In fact, in all conditions participants reported a mean of less than one source intrusion per list. The measure of source intrusion monitoring

utilised in Experiment 3.1 was the proportion of all generated source intrusions monitored correctly, for each participant in each Similarity condition. This meant that there were far greater numbers of source intrusions to analyse, and a larger data spread to detect differences.

Before the latency data can be discussed there is one very important caveat. Chi-squared goodness of fit analyses found that none of the ex-Gaussian fits were in fact a good fit to the bootstrapped data, which may explain many of the findings based on tau, mu and sigma. In addition, the effect sizes for the goodness of fit tests were not equal across the conditions. Therefore, the conditions differed in the reliability of their respective parameter estimates, complicating interpretation of the findings further. Overall, very few of the original predictions for this experiment were supported.

The first hypothesis (successful constrained search for both Similarity conditions) was only supported for the High-similarity trials. As expected, search set size as measured by tau was larger when recalling both sources than a single source, indicating successful search set size constriction in the High-similarity trials. This was supported by the behavioural data, as participants recalled fewer items in the single source condition than the both source condition. However, curiously, in the Low-similarity trials tau was larger for the Single source condition than Both sources. This is difficult to explain, as there is no reason to suggest why an instruction to recall less items, should cause an estimate of search set size to be larger when recalling a single source, than an instruction to recall the entire list. Indeed, the behavioural data suggest that this may be a misleading finding as participants recalled fewer items in the Low-similarity Single source condition than the Low-similarity Both sources condition, indicating successful constrained search.

In terms of the second hypothesis, it is also difficult to explain why the differences in tau between a single source and both sources were in opposite directions for the High and Low-similarity trials. In addition, it is unclear why search set size should be larger when recalling a single source in Low-similarity trials than High-similarity trials. One would expect that a less distinctive retrieval cue in High-similarity trials would cause more items to be searched than Low-similarity trials, whereas the reverse was found. Another issue is that this result is at odds with the recall data from the same experiment, which indicates that Similarity had no effect on search. Therefore, the difference in tau between recalling a single source and both sources should have been the same in both Similarity conditions and in the same direction. This is of course dependent upon the model being a good fit to the data, which we know was not the case in this instance.

From a theoretical point of view the estimates for tau contradict the predictions of the Source Monitoring Framework (Johnson et al. 1993), and to my knowledge, no model of retrieval would make these predictions. This also contradicts the EFR finding from Experiment 3.1 that there was no difference in constrained search between the two Similarity conditions. As stated in Chapter 3 It is possible that participants were able to find additional cues within the High-similarity trials, which assisted with discriminating the sources, for instance different accents for the male and female voices; however, this does not explain why estimated search set size was larger when recalling a single source than both sources in Low-similarity trials. Again it is likely that a poor ex-Gaussian fit is responsible for these findings which do not make theoretical sense.

A final puzzling finding was that estimates of tau indicated a larger set size in High similarity trials than Low-similarity trials when both sources were recalled. A less

distinctive retrieval cue for High-similarity trials may ultimately activate more items; however, recalling both sources could simply be executed using the current state of context, so there is no suggestion that Similarity should affect search set size when both sources are to be recalled. Again there is a caveat here, in that there is disagreement between the latency data and the recall data. There was good evidence that Similarity had no effect on total recall, indicating that the latency data may not be representative of participants' behaviour.

There was found to be no effect of Similarity or Recall instruction on the onset of recall. Although it was predicted that recall onset may have started sooner when both sources were recalled, this result is not unsurprising given that the prediction was not supported in the previous two experiments either. In Experiment 4.1 there was strong evidence for an effect in the opposite direction, and strong evidence for no difference in mu between a single source and both sources in Experiment 4.2. However, there was a main effect of Recall instruction on variability in onset of recall, with greater variability when both sources were recalled than a single source. Again this is surprising, as participants only need to locate and set a retrieval cue when a single source is recalled. For recall of both sources, the retrieval cue is the time of test context which is readily available at the start of the recall period. Therefore, one would expect greater variability in onset of recall for a single source. Overall, the latency data presented should be treated with great caution as they are based on poor fits, and in some cases are inconsistent with recall data from the same experiment.

4.5 - Impact of data aggregation on ex-Gaussian fits

On the whole, the curve fits presented in this chapter were significant departures from the ex-Gaussian, casting doubt over parameter estimates. It was predicted that this may be at least in part due to the data aggregation procedure used.

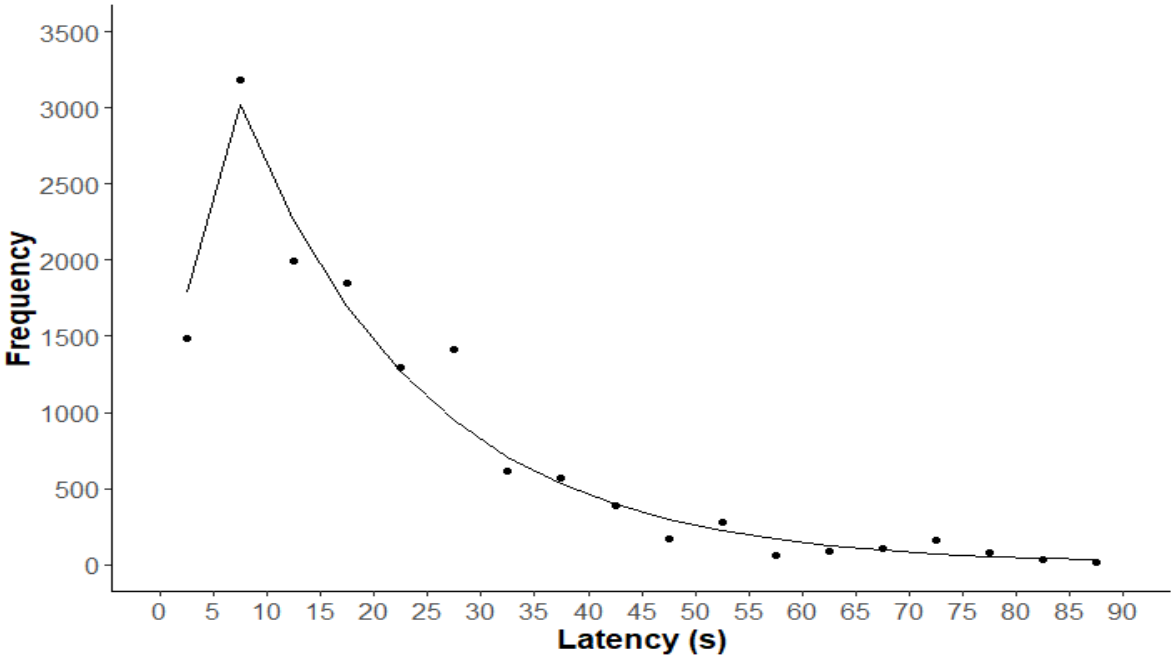
Recall in this experiment was too poor for individual participant fits; therefore, all latencies from all participants were combined, and then the data bootstrapped in order to derive ex-Gaussian parameter estimates. Of course, each participant would have initiated retrieval at different times, and retrieved items at different rates, potentially adding noise to the data once they were aggregated. In turn, this additional noise may have made a good ex-Gaussian fit less likely.

To investigate this possibility, thirty-six individual subject data were simulated, using the Both sources fit from Experiment 4.1 as a guide (guide fit). Total recall for each simulated participant was a random number of items drawn from a normal distribution, whose mean and standard deviation were the same as those for the recall data of the guide fit. Each simulated participant's recall latencies were generated by randomly sampling from an ex-Gaussian distribution. Values of mu, sigma and tau for each ex-Gaussian were randomly generated from a normal distribution, whose mean and standard deviation was equal to those of the bootstrapped parameter estimates for the guide fit. Recall latencies from all thirty-six simulated participants were combined into one dataset, and a best-fitting ex-Gaussian curve obtained using the fitting procedure detailed in section 4.2.1.5. Comparing the effect sizes of the chi-squared goodness of fit tests for the simulated data and the guide fit, will indicate the effect of aggregating data on the goodness of fit.

A chi-squared goodness of fit test revealed that as expected, the simulated bootstrapped data differed significantly from the best-fitting ex-Gaussian, $\chi^2(11) = 620.63$, $p<.001$, $w = 0.10$. See Figure 4.16 for the simulated data best-fitting ex-Gaussian. The corresponding effect size for the guide fit was $w = 0.12$. Therefore, we can say that data aggregation can account largely, but not completely, for a poor ex-Gaussian fit to the original Both sources data in Experiment 4.1.

Figure 4.16

Best-Fitting Ex-Gaussian for the Simulated Individual Subjects Data Based on Both Lists Fit from Experiment 4.1.



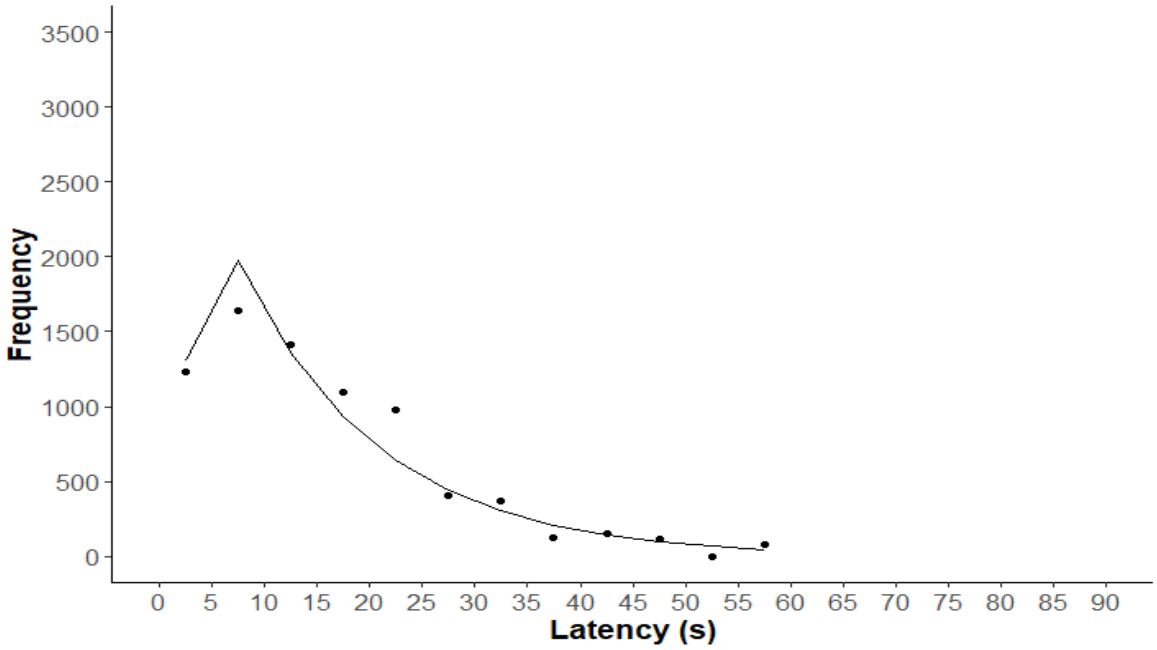
Note. Guide data was the Both sources fit from Experiment 4.1

This entire procedure was repeated using the High-similarity Single source fit of Experiment 4.3 as the guide fit, as this fit differed most strongly from the best-fitting ex-Gaussian of any of the fits presented in this chapter. See Figure 4.17 for the best fitting ex-Gaussian for this simulated data. By visually examining Figure 4.17, we can see that the best-fitting ex-Gaussian overestimates bins 1 and 2, and underestimates bins 4 and 5. This is the exact opposite of the guide fit (Figure 4.12), indicating that something other than data aggregation is responsible for the poor fit to the guide data. Further evidence can be obtained by comparing effect sizes. A Chi-squared goodness of fit test revealed that the simulated data significantly differed from the best-fitting ex-Gaussian, $\chi^2(11) = 423.87$, $p<.001$, $w = 0.12$. When compared with the effect size of the goodness of fit analysis for the High-similarity single source fit ($w = 0.20$), we see further evidence that data aggregation can only partly explain the deviation from the

best fitting ex-Gaussian. Other sources of noise in the underlying data may be largely responsible for the poor fit in this instance.

Figure 4.17

Best-Fitting Ex-Gaussian for Simulated Individual Subjects Data Based on High-Similarity Single Source Fit from Experiment 4.3.



Note. Guide data was the High-similarity Single source fit from Experiment 4.3.

4.6 - General discussion

The aim of the present chapter was to employ and test an alternative constrained search measure to EFR based on recall latencies, which does not suffer from selective reporting confounds. Thus, it served as a validation of EFR, in addition to an evaluation of a potential new method for measuring constrained search in the future. Replications of previous experiments presented in this thesis were conducted in order to directly compare constrained search across the methodologies. If EFR does not suffer from selective reporting confounds then the conclusions drawn from both methodologies should be largely the same.

Examination of the overt recall data across the two procedures were largely

reassuring in this regard. In all three Experiments reported in this chapter, it seems more likely than not that target availability was unaffected by procedure. There was also good evidence that total overt recall did not differ as a function of procedure in Experiments 4.1 and 4.3. In Experiment 4.2 the difference in overt recall was not significant; however, evidence for the null was only anecdotal. Taken together, it would seem that search processes were largely unaffected by experimental procedure, which is critical for the current treatment, as search processes are the primary concern of the present chapter. Therefore, we can conclude that these experimental procedures can be used in conjunction to investigate search processes.

If we take source intrusion output as a measure of source monitoring accuracy, then an interesting pattern emerges. In Experiments 4.2 and 4.3, where source was manipulated within a single list, there was sufficient evidence to assert that source intrusion output was higher in verbal-free recall than EFR, indicating superior source monitoring for the latter. This is likely due to EFR's explicit monitoring instruction for each generated item placing a greater emphasis on source monitoring quality. However, for List membership (Experiments 2.3 and 4.1) there was sufficient evidence that there was no difference in source intrusion output between the procedures. There could be two reasons for this. It may be the case that source intrusion monitoring is worse during verbal-free recall than EFR in List membership experiments, but there are insufficient source intrusions generated for this effect to be noticeable. Another explanation is that List membership as a source is fundamentally different in some way to Mixed-list sources. It may be the case that list is not a feature of an item that can be monitored by source; therefore, poorer monitoring of source features in verbal-free recall would have no effect on the output of source intrusions. Instead, monitoring in

such experiments could be accomplished using the same temporal context that drives search.

Before the latency data can be addressed, it is important to note the goodness of fit violations of the ex-Gaussian model fits to the data. This suggests the results should be taken with caution. Nevertheless, it is notable how consistent the parameter estimates of search set size from Experiments 4.1 and 4.2 supported the conclusions from EFR, in that search set size was smaller when recalling a single source than both sources, indicating constrained search. In addition, all curve fits match the general pattern of the data, despite the statistical departure from the ex-Gaussian, indicating that this is an appropriate distribution. A potential explanation for the statistically poor fits may be the way in which the data were aggregated. Overall recall in the experiments presented in this chapter was too low to perform individual subjects level fits. Instead, all latencies from all participants were combined, as if they originated from a single person. The ex-Gaussian fit, and the associated parameter estimates therefore are not representative of any given participant. For instance one individual's first recall could fall into the same bin as another individual's second recall. This could easily distort mu and sigma, and as a result affect tau. When this was formally investigated, it was found that data aggregation did affect the quality of fits, but only to a certain extent. Deviations from the ex-Gaussian are highly likely to be influenced also by other sources of noise in the underlying data.

Potential solutions to the data aggregation issue would be to increase list length, or to increase the number of trials for each condition. For instance, in Experiment 4.1, one could double the number of trials from three to six, requiring participants to recall List 1, List 2 and Both lists twice, then combining the latencies from each condition, hopefully achieving a large enough recall per condition to

perform a single subjects fit. In spite of this, the latency data will nonetheless be discussed, and some tentative conclusions drawn.

The first experiment presented was a replication of Experiment 2.3, where participants were required to study two lists of words and then selectively recall one of them. The latency analysis also required participants to recall both lists for one condition. Comparisons of the ex-Gaussian parameter tau across the three recall instructions suggested that participants searched more items when recalling Both lists than recalling either List 1 or List 2 in isolation indicating successful constrained search, thus supporting the EFR finding of Experiment 2.3. Furthermore, this seems a genuine finding as it is supported by the behavioural data from the same experiment.

One highly interesting finding was that estimated search set size was larger when recalling List 1 than List 2, suggesting that a retrieval cue for a prior event is more noisy than a cue for a more recent event. It is unlikely that a participant would be using time-of-test context to search for List 1, as one would expect a much larger value of mu for List 1 than List 2 which was not the case. Intriguingly this effect was not detected by EFR, as constrained search as assessed by aggregate scores and retrieval dynamics revealed no difference between recall of the 2 lists. If this is a genuine finding, which is important to note given the statistically poor ex-Gaussian fits, in addition to no such effect arising in the behavioural data then this raises questions regarding the sensitivity of accuracy measures such as EFR to detect more subtle effects of constrained search.

Experiment 4.2 was conducted as a latency measures replication of Experiment 3.2, with the aim of observing whether latency analysis could reproduce the EFR results that participants can constrain search in Mixed-lists. The prediction was also made that constrained search should be poorer for Mixed-lists than for List

membership. Importantly, estimated search set size was smaller for recall of a single source than both sources in Mixed-lists as expected. The difference in tau between a single source and both sources was also smaller than for List membership. It would appear that the overall difference between List membership and Mixed-lists was mostly driven by the Both sources condition. There was an effect of context on tau in the Single source condition however this was far smaller than the Both sources effect.

This points to Mixed-list sources being a weaker retrieval cue than temporal context for List membership. A weaker retrieval cue implies a less targeted search with more incorrect items being included in the search set, and reduced item availability when asked to recall all items. Broadly speaking this is reflected in both the latency analysis and behavioural data. There is a slight discrepancy whereby in the Single source condition, search set size is larger for Mixed-lists than List membership; however, recall is slightly lower for Mixed-lists than List membership when recalling a single source in the behavioural data. Despite this, we must remember that in verbal-free recall we do not have access to all generated source intrusions; many will be edited out by monitoring. In reality there may be far more source intrusions in the search set for Mixed-lists which are missed due to this, which will be detected by latency analysis.

Experiment 4.3 investigated the effect of Source Similarity on ability to constrain search in Mixed-lists. The equivalent EFR study, Experiment 3.1, demonstrated that Similarity had no effect at all on participants' ability to constrain search, as assessed by aggregate scores and retrieval dynamics. If EFR was sensitive enough to detect any significant differences in constrained search should they exist, then the same conclusions should be drawn from latency measures. However if EFR is

insufficiently sensitive, then latency analysis should have revealed a significant difference in constrained search between High and Low-similarity conditions.

The only finding that was expected was a larger estimated search set size for High similarity trials when recalling both sources compared with a single source, which supported Experiment 3.1. Unfortunately the majority of findings from Experiment 4.3 are difficult to explain as they contradict not only the hypotheses presented but also relevant theories relating to Similarity and source memory. It is very difficult to explain why an estimate of search set size should be significantly larger for recall of a single source than both sources in Low-similarity trials. In addition it is unknown why search set sizes should be larger for Low similarity than High-similarity trials for recall of a single source. This is completely contrary to the predictions of prominent source memory theories such as the Source Monitoring Framework, and there is no reason to suspect that a less distinctive retrieval cue should activate more source intrusions. One must note that none of the perplexing findings in the latency data are supported by the underlying recall data from the same experiment. In fact, the behavioural data largely support the equivalent EFR findings of no effect of Similarity on constrained search.

The present chapter demonstrates the promise of recall latency analysis as an assessment of constrained search. In Experiment 4.1, latency analysis seemed to be more sensitive to subtle differences in constrained search than EFR, when measuring differences between List 1 and List 2. It is also promising that despite the paucity of data from individual subjects and goodness of fit violations of the ex-Gaussian to the data, the expectation of smaller search set sizes for recall of a single source than both sources was met for Experiments 4.1 and 4.2. These findings can be seen as useful

first-pass evidence which should probably be followed up with a more data-rich replication using the methods suggested that might lend itself better to model-fitting.

Combining latency analysis and EFR will give researchers two separate measures of search accuracy which are complementary. The latter provides a rich assessment of exactly what has been searched and how search accuracy changes over time. The former has the highly useful property of not being influenced by selective reporting confounds which may affect search accuracy scores in EFR. Latency analysis can therefore provide further assurance for inferences drawn about search from EFR if the conclusions from the two methods agree. Therefore, future research into the role of constrained search (and monitoring using EFR) in recall accuracy should seek complementary evidence from both of these methodologies to draw conclusions, capitalising on the strengths of both.

Chapter 5: Computational modelling of output dynamics

5.1 - Introduction

Chapters 2 and 3 examined participants' ability to constrain search to a target source using a variety of contexts. In the majority of cases, participants preferentially generated targets over source intrusions at above chance level during a recall period, indicating an ability to selectively search for a target source. In addition, participants appear to nearly always retrieve a target at the start of a recall period, suggesting that they can easily locate an appropriate retrieval cue for the target source. After output position 1, the trend appears to be a steady decline in constrained search accuracy as a function of output position. At the same time, in almost all cases participants could monitor the output of their search with a high degree of accuracy. Target monitoring was near ceiling across the entire recall period, whereas source intrusion monitoring was generally at chance or below chance at output position 1, with a sharp increase thereafter.

The main issue with the interpretation of output dynamics data with respect to search accuracy, is that it is not possible to ascertain the psychological reason for the drop in search accuracy over time from the data alone. There are two possible explanations for this search accuracy trend. The first is that participants will inevitably find it more challenging to retrieve targets as the pool of novel targets decreases. It may be the case that a participant's ability to constrain search remains static throughout the recall period, and that the typical pattern of declining accuracy simply reflects the task becoming progressively more challenging as the number of remaining targets decreases. The second explanation is that participants cannot maintain the

target retrieval cue beyond output position 1; hence, the decline in search accuracy reflects deteriorating ability to constrain search. One way to separate constrained search ability from task difficulty is to build a computational model of the recall period which controls for the baseline rate of targets falling as recall progresses. In such a model, if output dynamics data from EFR experiments can be well described by a static probability of retrieving a target at each output position, it can be concluded that falling constrained search accuracy can be explained purely by a diminishing base rate of targets. Ultimately this would mean that participants' ability to search for targets does not worsen as the recall period progresses. If the data cannot be described by a static probability of recalling a target then this may indicate a deterioration in constrained search accuracy over time. The aim of this exercise was to see whether it is possible to build a simple retrieval model that can successfully reproduce the patterns of search accuracy data observed in EFR experiments presented in Chapter 2, in terms of output dynamics, targets and source intrusions generated, and dropout rate of lists by retrieval position.

5.2 - Model overview

The retrieval process is modelled as sampling with replacement (Polyn et al. 2009a; Raaijmakers & Shiffrin, 1981; Wixted & Rohrer, 1993). An item is selected from a pool of potential items. The retrieved item is then tagged as having been retrieved and replaced back in the search set. The pool of items available to retrieve is a subset of the total amount studied, with an equal probability of targets and non-targets being included. This is due to the target set being randomly defined by a noisy retrieval cue at test. Search accuracy essentially reflects the tendency to selectively retrieve targets over non-targets. Retrieval proceeds in this manner until a certain number of previously retrieved items are generated consecutively,

as described by Laming (2009). For simplicity, it is assumed that monitoring of repetitions is perfect, and there is no suppression of previously retrieved items from being retrieved again. All such repetitions are also assumed to be covert (not written down even if they were generated). This is feasible given that there were no specific instructions in the experiment as to what the participant should do in the case of retrieving a previously retrieved item. The model also does not attempt to account for primacy effects, or semantic, temporal (contiguity) or source clustering. It should also be noted that non-studied items, or items from a previous trial, are never output in this model. This is to reflect that such intrusions were very rare and were not considered when analysing the original data. For instance, if the fifth retrieval was a non-studied item, output position 5 would be assigned to the next studied item.

Three free parameters were employed. The first, p, is the probability of recalling a target at any given output position and stays static throughout. The second parameter, n, represents the number of targets in the search set. As a simplifying assumption, the same number of source intrusions are also contained within the search set, so the total number of items available is 2n. This assumption is based on the fact that the target set is defined by the retrieval cue. The final parameter, s, is the stopping rule for termination of recall. Recall terminates when s consecutive repetitions of any previously retrieved item occur.

As the data being modelled is the average of all participants in an experiment, n must be allowed to vary within a simulation, as it is virtually impossible for all participants in an experiment to have exactly the same memory capacity and to have selected a search set of identical size. Therefore, to integrate a degree of noise in the model's predictions the number of targets in the search set is treated as a unimodal distribution which approximates to normal, with n for that model iteration as the

mode. Thirteen simulations were run for each free parameter combination: Five where n=n, three where n=n+1 and n=n-1 and one where n=n+2 and n=n-2. The mean accuracy score for these thirteen simulations at each output position was then calculated. As the purpose of the simulation was to demonstrate what pattern of data this set of parameters would produce, this process was repeated one thousand times to gain the most stable estimates possible, without being too computationally expensive. The mean accuracy score at each output position from these one thousand runs was taken as the model's predictions for that parameter combination.

Given the variability in search set size incorporated into the model, not all generated outputs will be of equal length. It is possible therefore to predict the proportion of simulated data that contribute to each output position (dropout rate) for any given parameter combination. For each set of thirteen simulations, the dropout rate for output position x can be calculated by summing the number of outputs which have a length of x or greater, and dividing by 13. This process is run 1000 times. The model's prediction for the dropout rate at output position x for a given parameter combination, is the mean dropout rate at this output
position from these 1000 runs.

The recall period is modelled by first initialising a search set. The number of items in the search set is dependent on the value of n for the current model simulation. A repetition counter is also initialised, to be compared with s at each step in the recall period. Retrieval is modelled at each output position by generating a random number between 0 and 1. If this random value is less than or equal to p then a target is retrieved, if it exceeds p, then a source intrusion is retrieved. The retrieved item is tagged as having been retrieved and replaced in the search set. From output position 2 onward if a tagged item is retrieved, the repetition counter is incremented.

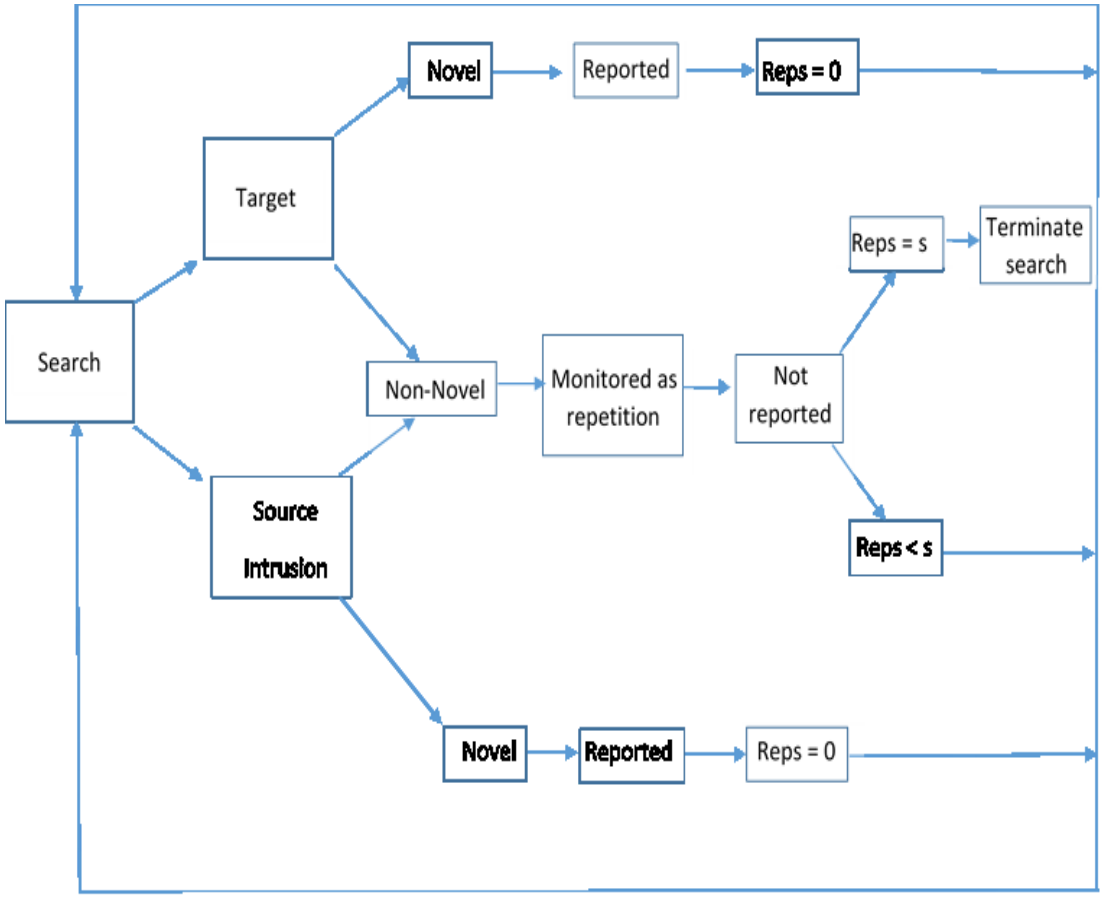
Whenever a novel item is retrieved, the repetition counter is reset to 0. Recall terminates when the repetition counter reaches s for the current simulation. If a simulated output terminates recall before four items have been retrieved, this output is discarded and replaced by another. See Figure 5.1 for a schematic representation of the

model.

It is also possible to obtain predictions for the mean number of targets and source intrusions generated for any given parameter combination. For each thirteen simulation block, the mean number of unique targets/source intrusions generated across these thirteen simulations is calculated. This process is then repeated one thousand times. The model's predictions for the number of targets/source intrusions generated for that parameter combination is the mean targets/source intrusions from these one thousand runs.

Figure 5.1

Schematic Representation of the Sampling With Replacement Model used to Simulate the Output Dynamics Data in Chapter 5.



Note. Reps = Repetitions, s = Stopping rule parameter.

5.3 - Impact of model parameters on predictions

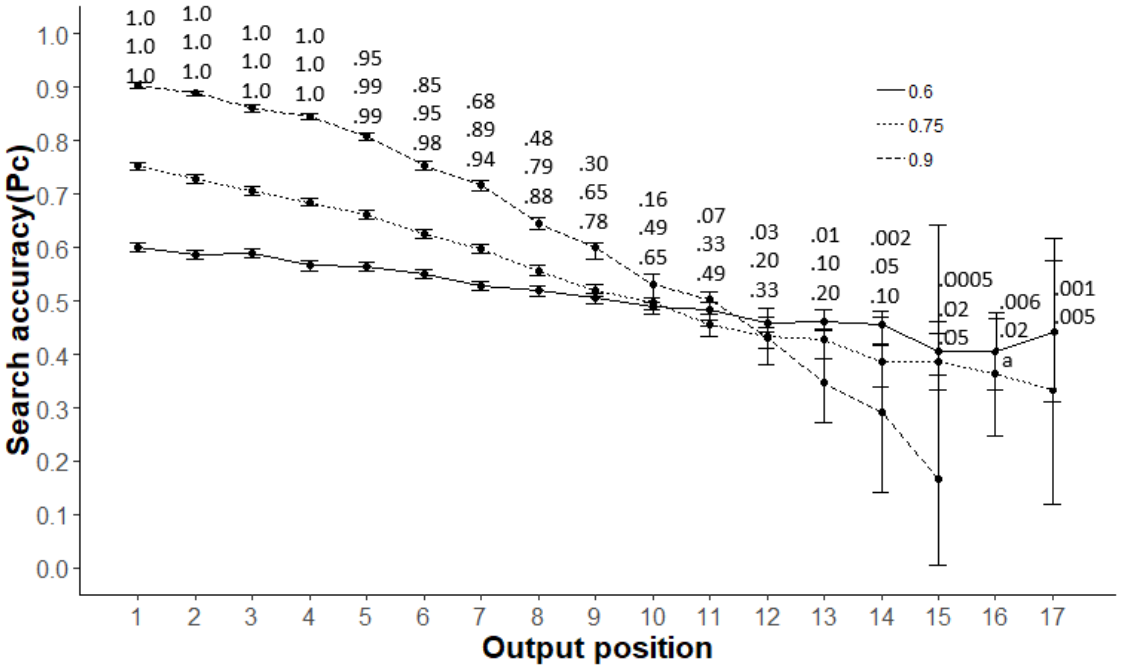
In order to observe how each parameter individually affected model predictions, three simulations were run. For each simulation, one of the three parameters was allowed to vary while the others were held constant. Curves were then plotted for each parameter depicting the relationship between search accuracy and output position, and list dropout rate and output position for three hypothetical values of that parameter. Each simulation contained 100,000 iterations to achieve smooth curves. These hypothetical output dynamics and dropout rate curves are depicted in Figures 5.2 - 5.4.

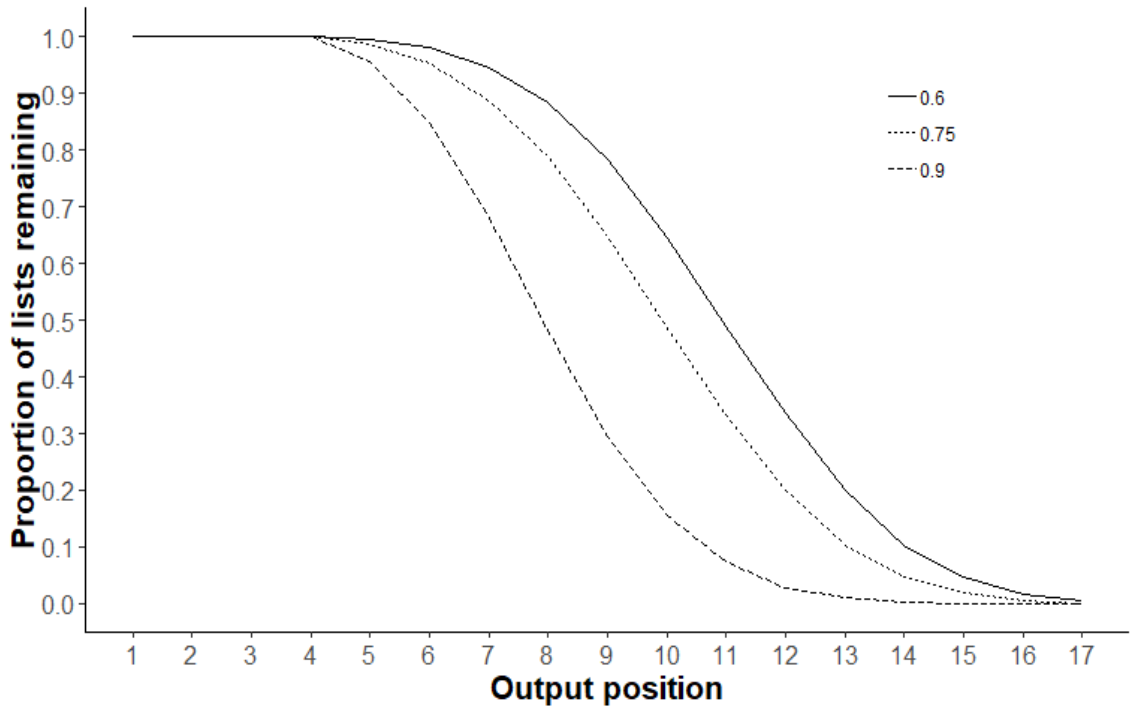
By examining the top panel of Figure 5.2, we can see that at output position 1,

search accuracy reflects the value of target recall probability (p) set for that simulation. As the recall period progresses, the decline in search accuracy gets steeper with increasing p. The higher the value of p, the less likely it is that novel targets will still be available later in the recall period, so search accuracy drops more sharply. As can be seen on the bottom panel of Figure 5.2, p is likely to also have an effect on the dropout rate. With higher values of p, the pool of novel targets will reduce more quickly. This also increases the probability of retrieving a previously generated target, which will increment the repetition counter more quickly, leading to simulated recall periods terminating earlier. Finally, a higher value of p should have a positive effect on the number of simulated targets retrieved, and a negative effect on the number of intrusions retrieved.

Figure 5.2

Hypothetical Curves for the Effect of Target Recall Probability on Search Accuracy (Top) and Dropout Rate (Bottom).





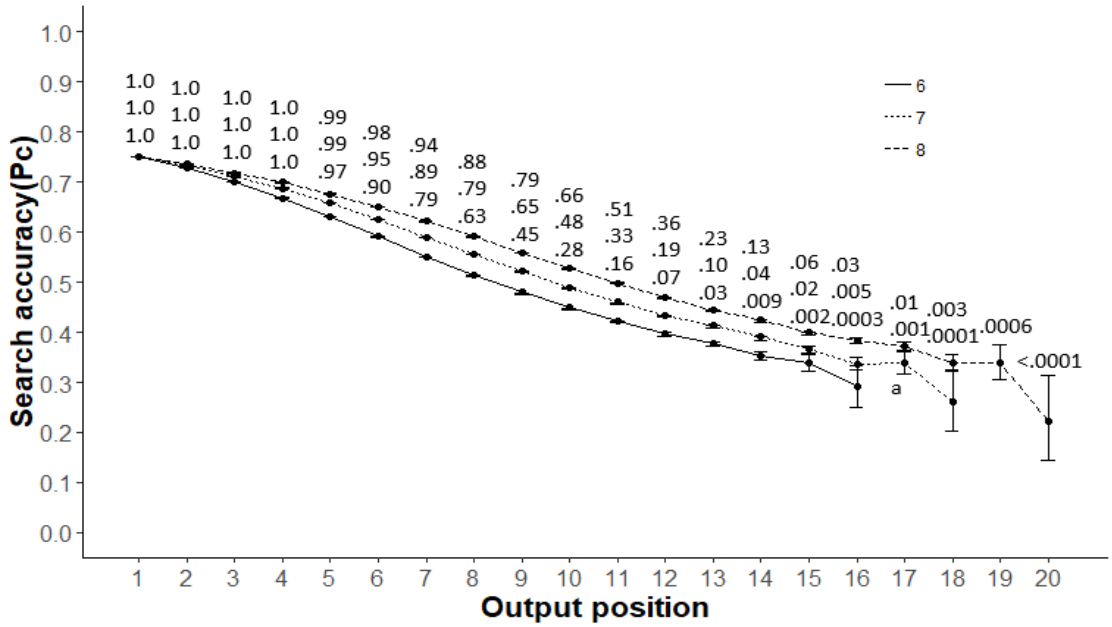
Note. For this simulation n=8 and s=5. It should be noted that n is the mode of a unimodal distribution of target numbers. The maximum number of targets is in fact n+2. Hence, why the dropout rate is non-zero beyond output position 16 for p=0.6. Error bars represent 95% confidence intervals for each output position. Digits above data points show proportion of total simulated data contributing to each output position. Top, middle and bottom digits represent p=0.9, p=0.75 and p=0.6 respectively.

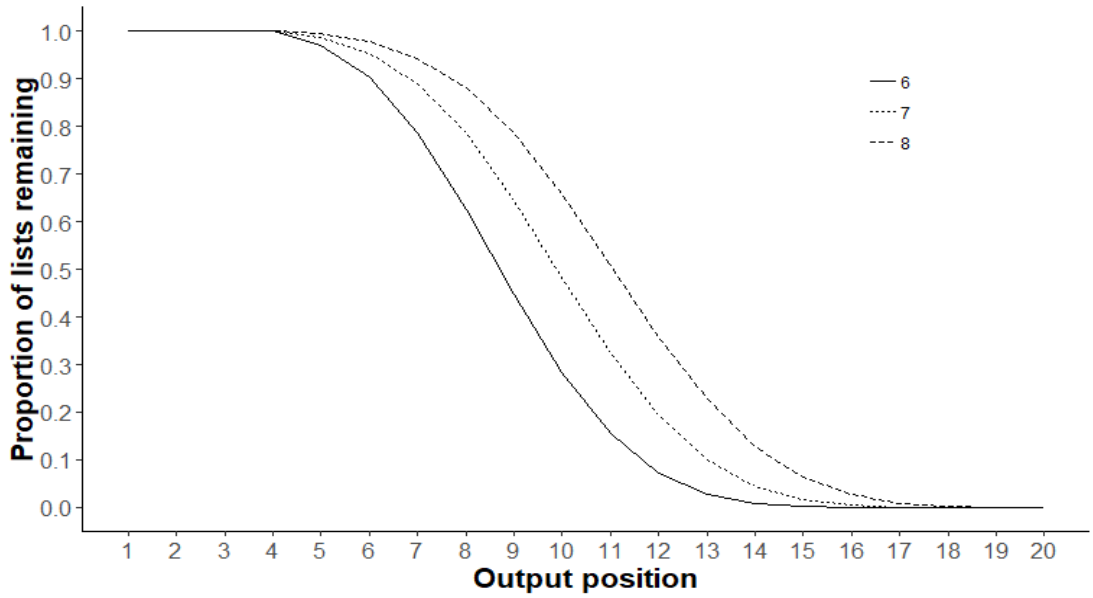
[a] For output positions 16 and 17, the top digit represents p=0.75 and the bottom digit represents p=0.6

Figure 5.3 describes the effect of the modal number of targets in the search set (n) on search accuracy and dropout rate. The first effect of note is the effect of n on rate of search accuracy decline, with higher values of n yielding a shallower curve. At any given output position, there are more novel targets available to retrieve with higher values of n than lower values. In terms of the dropout rate, this should be more shallow as n increases, because a higher n leads to a greater probability of a novel item being retrieved than a lower n. In terms of targets and intrusions retrieved, both should increase with n.

Figure 5.3

Hypothetical Curves for the Effect of Modal Number of Targets (n) on Search Accuracy (Top) and Dropout Rate (Bottom).





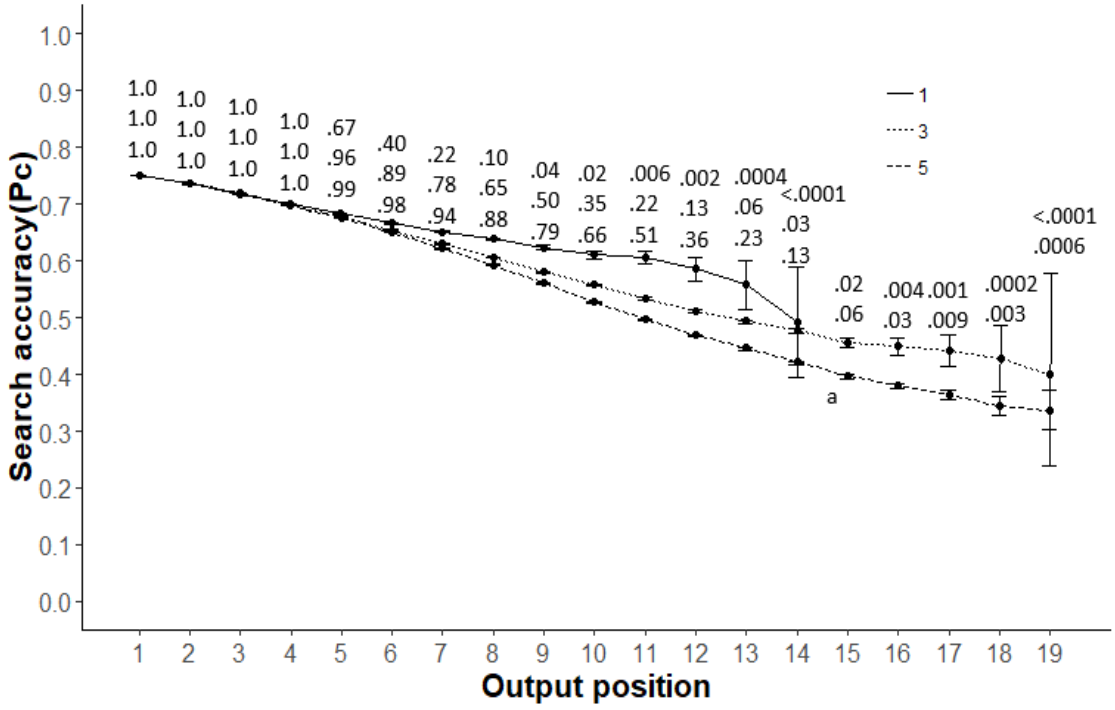
Note. For these simulation p=0.75 and s=5. Error bars represent 95% confidence intervals for each output position. It should be noted that n is the mode of a unimodal distribution of target numbers. The maximum number of targets is in fact n+2. Hence why the data extends beyond output position 2n for each value of n. Digits above data points show the proportion of simulated data contributing to each output position. Top, middle and bottom digits represent n=8, n=7 and n=6 respectively.

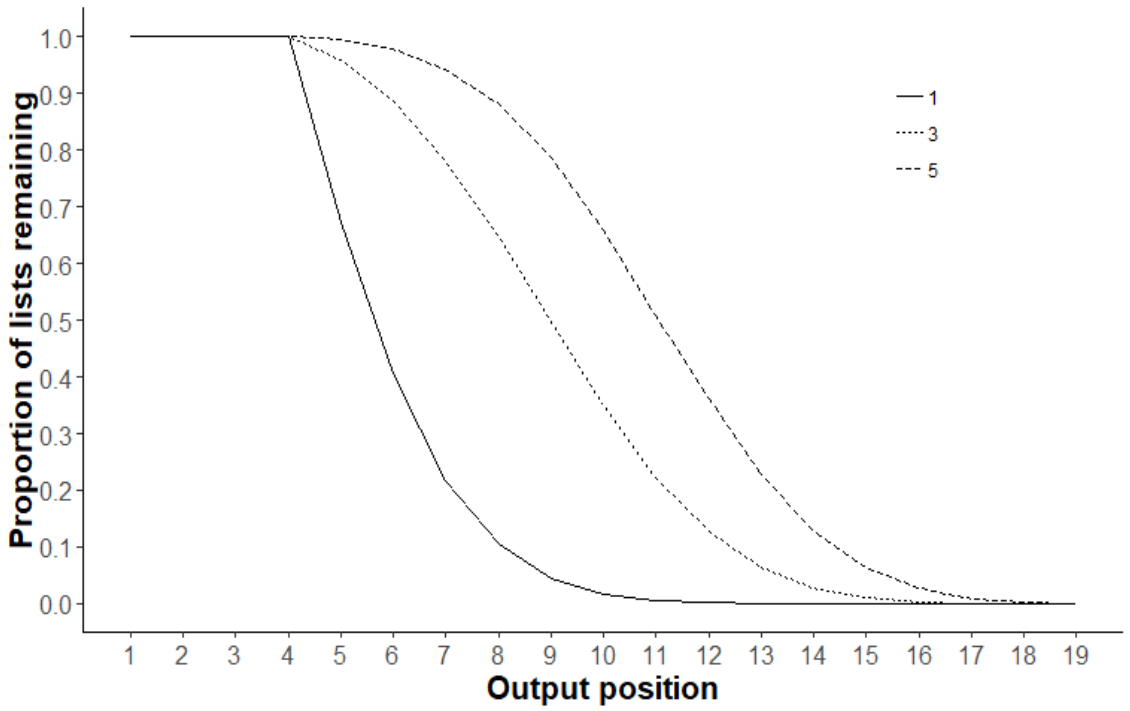
[a] For output positions 17 and 18, top and bottom digits represent n=8 and n=7 respectively

Figure 5.4 shows that the effect of s on search accuracy appears to be shallower decline for lower s values. The bottom panel shows that higher s values will lead to a shallower dropout rate than lower s values as expected, given that this parameter sets the permissible number of retrieval failures (repetitions of previously retrieved items) before recall is terminated. Finally, a higher value of s will lead to greater numbers of targets and intrusions retrieved.

Figure 5.4

Hypothetical Curves for Effect of Stopping Rule (s) on Search Accuracy (Top) and Dropout Rate (Bottom).





Note. For these simulations, p=0.75, n=8. It should be noted that n is the mode of a unimodal distribution of target numbers. The maximum number of targets is in fact n+2. Hence why the data extends beyond output position 16 for s=3 and s=5. Error bars represent 95% confidence intervals for each output position. Digits above data points represent the proportion of simulated data contributing to each output position. Top, middle and bottom row represent s=5, s=3 and s=1 respectively

[a] For output positions 15 -19, top digit represents s=3 and bottom digit represents s=5.

5.4 - Model fitting

Simulations of the output dynamics data from Experiment 2.3 were conducted. A good fit of the model to the output dynamics, targets and source intrusions and dropout rate data would indicate that the typical pattern of deteriorating search accuracy over time can be attributed to a falling base rate of targets over this same period.

A predefined set of parameter values were initialised prior to running the simulations. Values of p ranged from 0.5 to 1 in steps of 0.01, values of n ranged from 6 to 8 and values of s ranged from 1 to 10. All outputs where the participant had recalled more than three items were included in the simulations. Note that the

parameter values themselves were not used to draw conclusions about the data. The idea was to simply explore the full range of parameter values, and to observe if the best fitting combination gave an accurate representation of the data.

Parameter estimation was accomplished by the method of Least Squares Estimation. Each simulation with a particular combination of parameter values produced a search accuracy curve to be compared with that from the real data. Finding the best fitting model was complicated by the fact that after output position 4, progressively fewer lists contributed to each output position. Therefore, the data becomes increasingly noisy as the recall period progresses. Accordingly, a least squared estimate was devised, which weights the discrepancy between the search accuracy data and the model's predictions at each output position by the number of lists contributing to each output position. This was termed the Weighted Root Mean Squared Deviation (WRMSD). At each output position, the discrepancy between the model's predictions for search accuracy and the real data is multiplied by the number of lists contributing to that output position, and squared. This is then divided by the total number of weighted data points. The WRMSD is the square root of the mean of this value across all output positions. The result is a least squared estimate that is much more sensitive to deviations from the output dynamics data at earlier output positions than later ones. In short, it is more important for the model's predictions to be closer to the search accuracy data early in the recall period than later. The best fitting parameter combination was that which resulted in the lowest WRMSD.

Once the best fitting model was found, its goodness of fit was assessed by visually inspecting the observed plot of the data, comparing the curves of the model's predictions of accuracy by output position with the observed data. Due to the small number of output positions in the underlying data, goodness of fit could not be

assessed by parametric tests. Other than the search dynamics, a number of measures of goodness of fit were then assessed, such as the list dropout rate by output position. Visual inspection of data vs model dropout plots were employed to assess goodness of fit on this measure. The final measure of goodness of fit was whether the model could accurately predict the mean number of targets and source intrusions generated in that experiment. The modelling approach was iterative. Initially, the simplest form of the model was fitted (perfect repetition monitoring) and assessed for goodness of fit. Then a second iteration was tested with modified assumptions (imperfect repetition monitoring). Its performance was compared with that of the first iteration to determine if this improved the fit.

5.5 - Model iteration 1

The following simulation refers to the output dynamics data obtained from Experiment 2.3, which was the first experiment to use the EFR methodology. To recap, in this experiment, participants were given EFR instructions to recall one of two lists separated by a thirty-second delay in either the Visual or Auditory modality. The main point of interest was whether the model can successfully predict the data collapsed across List membership and Modality in terms of search dynamics, dropout rate and numbers of targets and intrusions.

This simulation examined whether a static value of p for the whole recall period could adequately describe the data. If this is the case, then one can conclude that it is not necessary to assume declining search efficiency with output position. Best fitting parameter estimates are presented in Table 5.1. When examining Figure 5.5, we can see that the model's predictions generally predict the pattern of observed search dynamics quite well with the exception of a slight underestimation at output position 1 and overestimation at output position 4. There was also a fairly sizeable

overestimation late on in the recall period where the underlying data were inherently noisy.

Table 5.1

Best Fitting Parameter Estimates and Deviation Scores for the First (Perfect Repetition Monitoring) and Second (Imperfect Repetition Monitoring) Iterations of the Model.

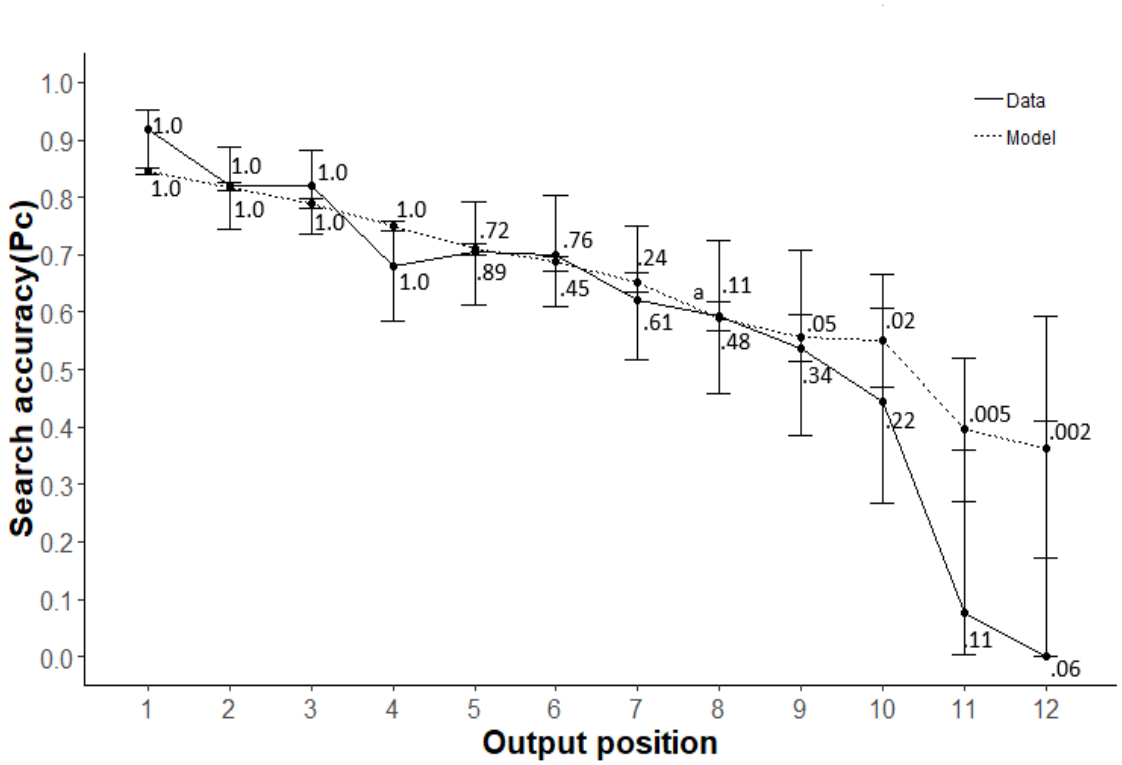
Parameter	Model iteration 1	Model iteration 2
p	0.84	0.84
n	6	6
s	2	3
repm	----	1
WRMSD	0.14	0.13

Note. WRMSD = Weighted Root Mean Squared Deviation

However, by examining the dropout rate as expressed in Figure 5.6, one can see that after the first four output positions where the dropout rate is constant, the model's predicted dropout rate is far more severe than the underlying data. It is worth noting the high degree of error in the observed accuracy data in later output positions.

Figure 5.5

Modelling of Search dynamics data from Experiment 2.3 collapsed across lists and modalities.

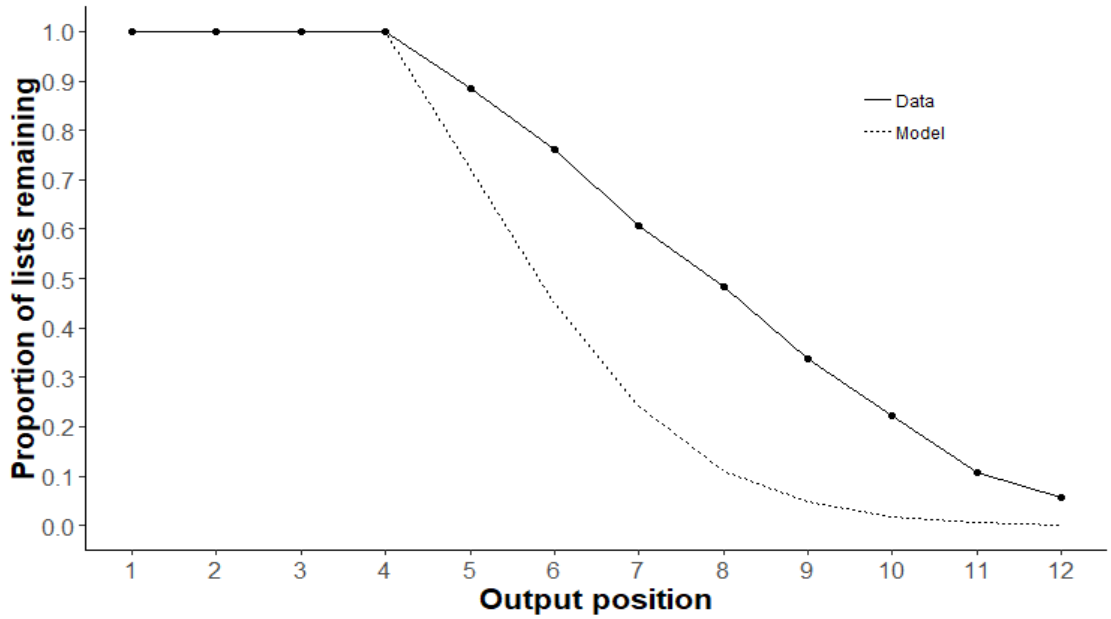


Note. Error bars represent 95% confidence intervals for the model predictions and original data at each output position. Digits above/below each output position represent the proportion of total real data and simulated data respectively, contributing to each output position.

[a] Proportion of real data = .48. Proportion of simulated data = .11

Figure 5.6

Observed and Predicted Dropout Rates for Experiment 2.3 Data, Collapsed Across Modalities and List Membership.



In addition, as can be seen from Table 5.2, the model appears to underestimate the total number of targets and source intrusions retrieved. This is indicative of premature termination of the recall period when compared with the data. One potential explanation is that the model assumes two things regarding repetitions. The first is that repetitions are noted but not recorded. The second is that repetition monitoring is perfect. These assumptions are somewhat unrealistic in reality, given that repetitions are occasionally observed in EFR outputs (Unsworth et al., 2010). It is more likely that some generated repetitions are mistaken as novel items and are overtly output. Given that the stopping rule is based on retrieval of consecutive repetitions, it is possible that imperfect repetition monitoring may extend recall periods due to incorrectly monitoring some repetitions as novel items. This idea is explored in a second model iteration detailed in the next section.

Table 5.2

Predicted and Observed Numbers of Targets and Source Intrusions for the First (Perfect Repetition Monitoring) and second (Imperfect Repetition Monitoring) Iterations of the Model.

Measure	Model iteration 1	Model iteration 2
Predicted Targets	4.29	4.70
Observed Targets	5.16	5.16
Predicted Source intrusions	1.31	1.60
Observed source intrusions	2.06	2.06

5.6 - Model iteration 2

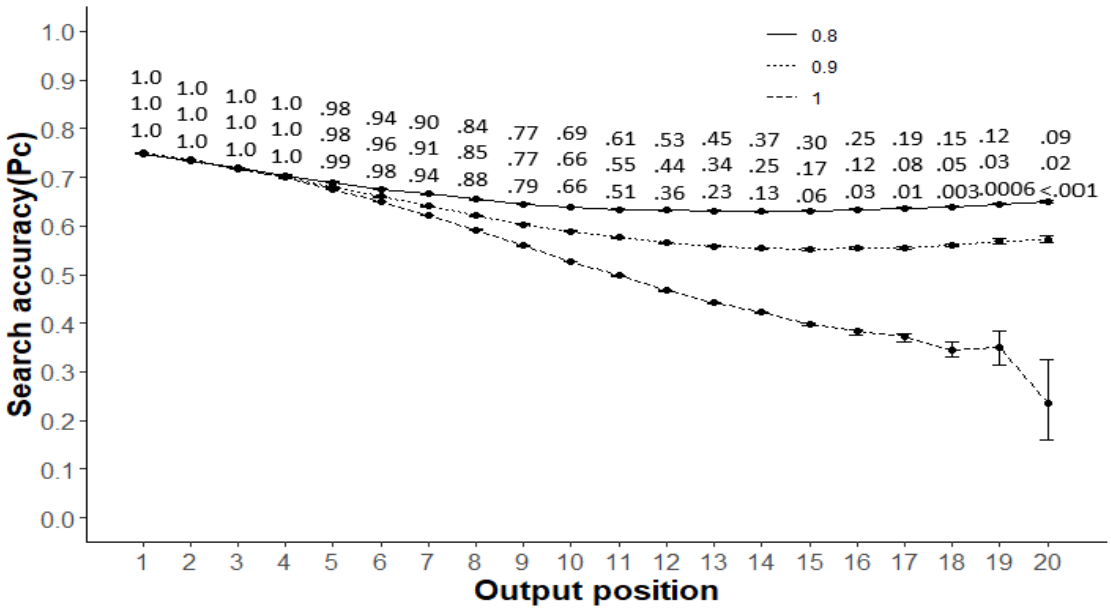
To investigate the role of repetition monitoring efficiency on model predictions, an additional free parameter repm was added, which controlled for the accuracy of repetition monitoring. Specifically the value of repm represents the probability that a repetition will be monitored as a repetition, and a novel item will be monitored as a novel item. If a repetition is incorrectly monitored as a novel item the repetition counter will be reset to 0, leading to an extension of the recall period. The second condition was included as it is of course possible that a participant may believe that they have previously retrieved an item that was in fact novel. repm remains static across the entire recall period. The impact of repm on search accuracy and dropout rate is depicted in Figure 5.7.

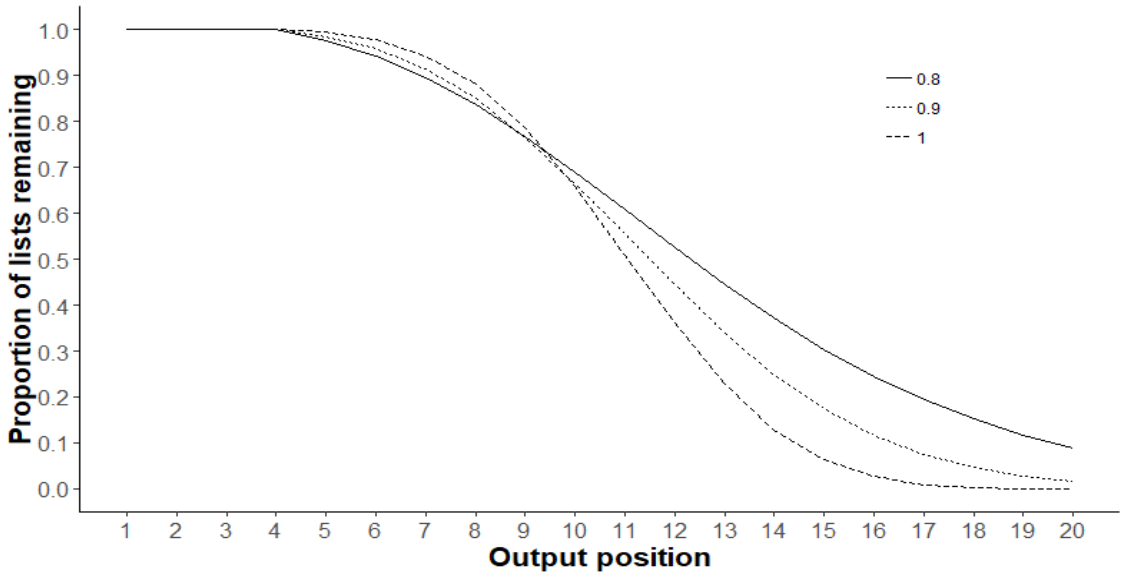
The top panel of Figure 5.7 demonstrates a sharper decrease in search accuracy as values of repm increase. As repetition monitoring accuracy gets poorer, more repetitions will be mistaken as novel items, which will reset the repetition counter more frequently. This provides more opportunities to retrieve novel targets further into the recall period. The effects of repm on the dropout rate are related to the likelihood of a repetition occurring at different points in the recall period. Early in the recall period repetitions are less common; therefore, repetition monitoring errors are

most likely to be those where novel items are incorrectly monitored as repetitions. In this situation, the error (repetition) counter is more likely to be incremented when repetition monitoring is poorer; hence, why less data drops out with higher values of repm. As the recall period progresses, repetitions become increasingly likely, such that repetition monitoring errors are predominantly those where repetitions are mistaken as novel items. In this case, poorer repetition monitoring will cause the repetition counter to be reset to 0 more often, leading to a more shallow dropout.

Figure 5.7

Effect of Repetition Monitoring Accuracy (repm) on Search Accuracy (Top) and Dropout Rate (Bottom).





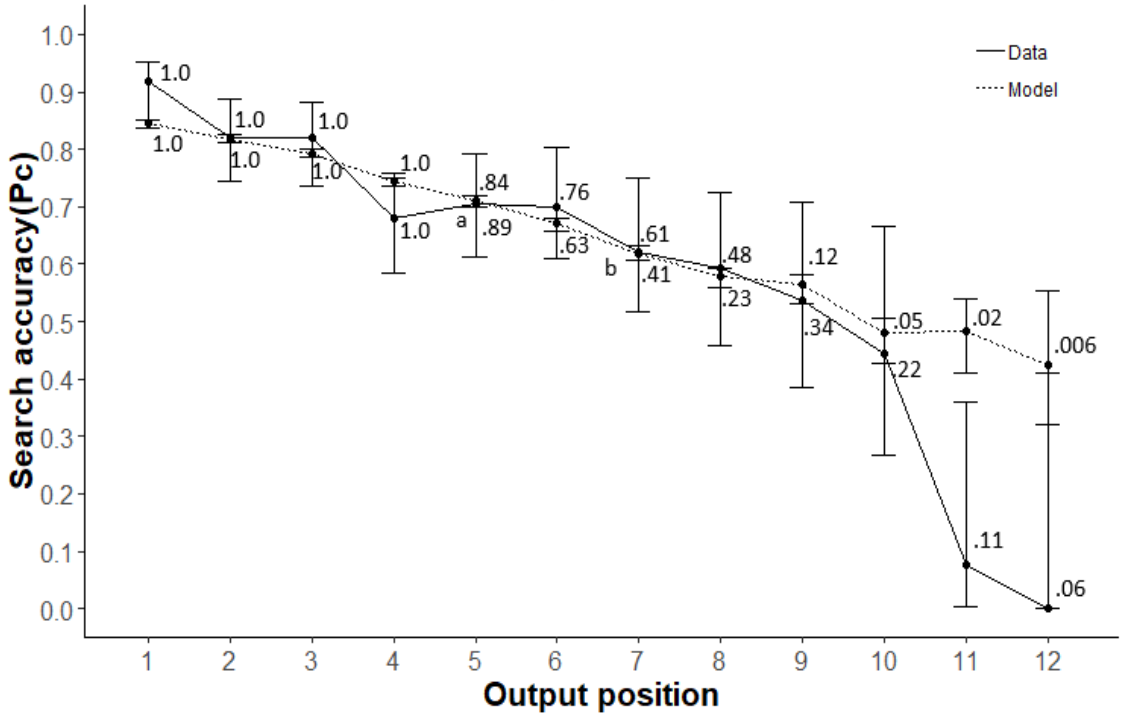
Note. Error bars represent 95% confidence intervals for each output position. p=0.75, n=8. It should be noted that n is the mode of a unimodal distribution of target numbers. The maximum number of targets is in fact n+2. Hence why the dropout rate is non-zero beyond output position 16. Digits above data points represent the proportion of simulated data contributing to each output position. Top, middle and bottom digits represent repm = 0.8, repm = 0.9 and repm = 1.0 respectively.

Fits of the second iteration of the model to the data are presented in Figure 5.8. Again, it seems a similar pattern emerges to the first model iteration presented in Figure 5.5, with an underestimation of the accuracy data at output position 1 and an overestimation at output position 4. When we compare the performances of the two model iterations by examining their relative WRMSD scores (see Table 5.1) we can see that the second iteration (imperfect repetition monitoring) is only marginally superior in terms of quality of fit to the search dynamics data. Table 5.2 shows that there was also a small improvement for model iteration 2 in terms of the number of targets and source intrusions generated although these metrics were still underestimated compared with the observed data. When comparing Figures 5.6 and 5.9 it would seem that model iteration 2 was also a better fit to the dropout data than the first iteration; however, the predicted dropout was still too steep.

Interestingly though, this improved fit to the dropout rate and number of targets and source intrusions retrieved for model iteration 2 cannot be attributed to imperfect repetition monitoring. As Table 5.1 shows, the best fitting parameter value for repm was 1, or perfect repetition monitoring. Given that p and n were identical across the 2 model iterations, the improved fit to the dropout rate and recall metrics can only be due to a more liberal stopping rule (increased s). This does pose problems regarding the consistency of parameter estimates. If we remove repm from the equation given that the best fitting value for this parameter was 1, it would be expected that values for the remaining three parameters would be identical for the two iterations as the assumptions regarding retrieval are the same. However this was not the case, suggesting that every time the model is run, different best fitting parameter values could be observed, which calls into question the model's reliability. This will be investigated further in the next section.

Figure 5.8

Model Iteration 2 Fit to Search Dynamics Data from Experiment 2.3 Collapsed Across Modalities and List Membership.



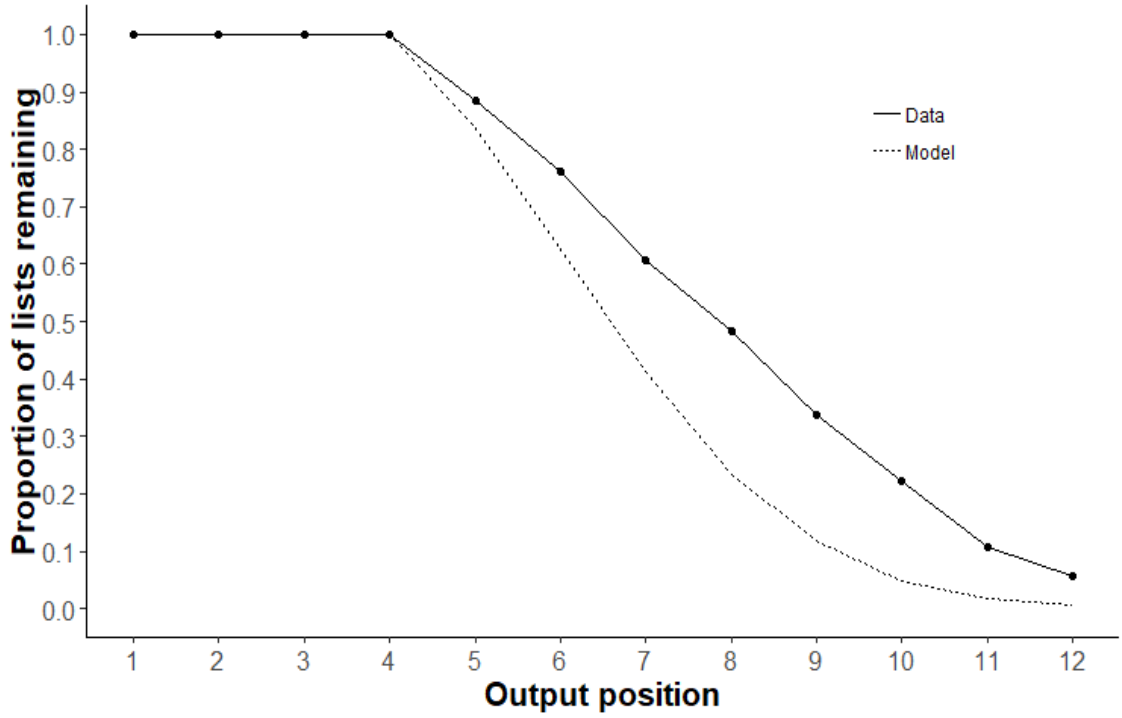
Note. Error bars represent 95% confidence intervals for the model predictions and original data at each output position. Digits above/below data points represent the proportion of total real and simulated data respectively, that contribute to each output position.

[a] Proportion of real data = .89. Proportion of simulated data = .84.

[b] Proportion of real data = .61. Proportion of simulated data = .41.

Figure 5.9

Model Iteration 2 Predicted and Observed Dropout Rates for Fully Collapsed Experiment 2.3 Data.



5.7 - Parameter Recovery Exercise

The fact that the model returned different parameter values when fitting the same dataset on two separate runs begs the question of whether the model can be defined. To test this, a parameter recovery exercise was conducted where four datasets were simulated with known parameter combinations spread across the parameter space. The same model fitting procedure as described in section 5.4 was applied to the simulated recall outputs. If the model can be reliably defined, then the recovered parameter values from the fitting procedure should be identical to those used to create the original simulated outputs. See Table 5.3 for the known and recovered parameters from this exercise.

Table 5.3

Known and Recovered Parameters from Parameter Recovery Exercise

Comb	p		n		s		repm	
	Known	Rec	Known	Rec	Known	Rec	Known	Rec
1	0.90	0.90	6	6	2	3	0.80	0.77
2	0.80	0.80	7	7	3	2	0.85	0.88
3	0.70	0.70	8	8	5	4	0.90	0.92
4	0.60	0.60	9	9	1	2	0.95	0.89

Note. Rec = Recovered, Comb = Combination

By examining Table 5.3 a clear pattern emerges. Parameters p and n appear to be very stable, as recovered values for these parameters perfectly matched the known parameter values for each combination. However, it also appears that parameters s and repm are very unstable, as none of the recovered parameter values matched the known parameters. The likely reason for this is that p and n seem to strongly affect accuracy between output positions 1 and 4 (see Figures 5.2 and 5.3), where the least squared estimate (WRMSD) is most heavily weighted. Therefore, tested parameter values which deviate only slightly from the known parameter values will still have a sizeable effect on WRMSD, making it likely that the recovered parameter values will match known values. Parameters s and repm appear to affect WRMSD much more subtly, as their effect on accuracy tends to be observed much later in the recall period (see Figures 5.4 and 5.7). Therefore, there is more scope for the best-fitting values of these parameters to deviate from the known parameters. As a result, it would seem that the model cannot reliably defined. Given this, one should be very cautious when attempting to draw conclusions from this modelling exercise. This issue would likely be resolved by increasing the sample size for each simulation to obtain more stable estimates.

5.8 - General discussion

In Chapters 2 and 3 we saw that output dynamics can offer an interesting insight into how constrained search accuracy progresses over time. However, computational modelling techniques allow us to shed light on theoretical questions which cannot be answered by output dynamics alone. The primary aim of this exercise was to see whether the output dynamics data from Chapter 2 could be explained by a simple sampling with replacement retrieval model, initially with three parameters: The probability of retrieving a target (p), which remained static throughout the recall period, the number of targets in the search set (n) and a stopping rule based on the number of consecutive repeated retrievals (s). For the model to be deemed a good fit to the data, it was also necessary that it could predict other aspects of the data such as dropout rate and number of targets and source intrusions retrieved.

The first iteration of the model assumed that repetition monitoring was perfect, and that repetitions were not recorded even when searched. This was found to be a reasonable fit to the search dynamics data although the number of targets and source intrusions retrieved were underestimated, and the slope of the dropout rate was too steep. This was assumed to be as a result of the recall period terminating prematurely, possibly due to repetition monitoring being perfectly accurate.

As such, a second iteration of the model was tested. This employed an additional parameter (repm) which controlled for the accuracy of repetition monitoring. This allowed for errors to be made, such as mistaking repetitions for novel items and vice versa. On the whole, this free parameter had no effect on the quality of the fit to the search dynamics data, given that the best fitting parameter value corresponded to perfect repetition monitoring. The improvement of the fit to the recall metrics and dropout rate was determined to be due to a more liberal stopping

rule.

It was however noted that parameter estimates were unreliable and would not necessarily replicate across simulations. This is probably related to the sample size and computational expense. Unfortunately the simulations presented were limited by technological constraints. The total time taken to run an entire simulation of model iteration 2 with one thousand samples for each parameter combination was around thirty-six hours. Increasing the sample size would have been impractical. One solution would be to increase the sample size, yet restrict the potential parameter estimates to more realistic values. This could be informed by the hypothetical accuracy curves presented in Figures 5.2-5.4 and 5.7.

Despite this, from the present modelling effort we can conclude that a simple explanation based on a reducing target pool with static target recall probability throughout the recall period, is unlikely to completely account for the pattern of search dynamics observed in Experiment 2.3, and that there must be other factors at play.

One such factor which was not addressed here but could be explored is the potential influence of selective reporting. The model at present assumes that all generated source intrusions are recorded. However, as previously stated, the main criticism of EFR is the potential for participants to not record generated source intrusions in an attempt to artificially improve their search accuracy scores. A free parameter could be implemented which controls for the frequency of such failures to record source intrusions. Doing this may give an insight into the extent to which search accuracy is artificially inflated by selective reporting in the data presented in this thesis.

Inevitably this will be affected by the accuracy of source monitoring. Target

monitoring can be disregarded as targets were very rarely mistaken as source intrusions in the EFR experiments presented in this thesis. Therefore, a free parameter controlling the probability of making a source intrusion monitoring error would be implemented. However, this is complicated by the fact that source intrusion monitoring accuracy is not static throughout the recall period. There would appear to be source neglect at output position 1, manifesting in the majority of source intrusions being monitored incorrectly. Source intrusion monitoring appears to engage at output position 2 and increase steadily to ceiling thereafter. In order to capture this, a free parameter controlling for source intrusion monitoring would have to take into account source neglect and increasing source intrusion monitoring from output position 2 onwards, in order to correctly test the influence of source monitoring on the search dynamics data presented.

There is of course the possibility that search efficiency does diminish throughout the recall period. Such a model would allow for the probability of retrieving a target (p) to decrease with output position. The rate of this decline would be controlled by a decay parameter. This would have the effect of reducing search accuracy predictions at each output position. It would also likely extend the recall period by allowing more source intrusions to be generated instead of repeatedly searching previously retrieved targets.

One final point to make is that the objective of this exercise was to build a basic model that would allow testing of a simple explanation for the observed data, rather than to construct a comprehensive model of retrieval. Therefore, the presented model did not attempt to account for recall phenomena such as clustering (temporal and source), and primacy and recency effects. In addition, while sampling with replacement is frequently employed as a viable method of describing retrieval

processes (Bousfield & Sedgewick, 1944; Wixted & Rohrer, 1993), this mechanism does lend itself to outputting an unrealistically large number of previously retrieved items (repetitions). This is not to say that sampling with replacement is inappropriate, as more contemporary models do continue to describe retrieval in a similar fashion (Polyn et al. 2009a). However, a provision is made whereby previously retrieved items are suppressed in some way to reduce the likelihood of them being retrieved again. If one were interested in gaining a more in depth insight into the processes at work beyond testing a very simple hypothesis for the data pattern, a more comprehensive model could be constructed which does account for the aforementioned recall phenomena, and restricts repetitions; thus, yielding more realistic recall outputs.

Chapter 6: General discussion

This thesis explored the joint contributions of constrained search and source monitoring to the control of recall accuracy. The basic premise being that when required to recall a specific event, we attempt to reinstate the context of the original event using source cues, to constrain memory search to correct information. However, this is fallible, so every piece of information that comes to mind is monitored for correctness. If the information passes a correctness criterion, then it is overtly reported. If it does not, then the information is withheld (Goldsmith, 2016).

Due to the effectiveness of the monitoring mechanism, standard-free recall is inappropriate to study this, as many of the errors generated during search are edited out and are not reported; therefore, we have little insight into the accuracy of constrained search. Chapter 2 set out to develop and test a suitable paradigm that could simultaneously measure constrained search and monitoring. Externalised-Free Recall (EFR) is ideal for this purpose as participants are instructed to report correct and incorrect information that comes to mind (Bousfield & Rosner, 1970; Kahana et al. 2005). In addition, every time a participant reports an item which they believe was incorrect, they must press a button on a computer keyboard. Presuming that the participant is reporting errors as well as correct information, this gives an accurate representation of what the participant has searched, and how accurately they have monitored intrusions. The present thesis employed a modified version of EFR, requiring participants to make monitoring judgments on each item reported. This was to avoid a potential confound with the keypress methodology, whereby the

participant forgets to press the key to indicate that the retrieval is an intrusion.

In Chapter 2, source was defined as List membership. This was ideal to test the new paradigm as participants are known to be able to specifically isolate a given list in the presence of another (Jang & Huber, 2008; Unsworth et al. 2013). The key questions were whether the paradigm reflects retrieval processes observed in standard-free recall, was the paradigm sufficiently sensitive to detect successful constrained search, and; furthermore, how does this relate to and react with source monitoring to control recall accuracy?

The first methodological issue to address was whether recall outputs from an EFR experiment are in fact representative of known and well established retrieval processes that occur during recall. Standard-free recall (SFR) is edited recall, whereby the participant does not report items that are monitored as incorrect. For EFR reporting is unedited, as the participants are required to report correct and incorrect items. An accurate assessment of search therefore is reliant upon the participant reporting incorrect items, which is a concern. Additionally, the requirement to monitor each reported item may affect the organisation of recall outputs by source, due to the increased salience of source in EFR. One might expect a greater degree of organisation (clustering) by source in EFR than in SFR. Three experiments were run to examine the impact of unedited reporting and monitoring instructions on quantity of recall, accuracy of overt recall, and organisation of recall. The first was a standard-free recall experiment with no monitoring instruction, the second, a standard-free recall experiment with a monitoring instruction, and finally EFR. Reassuringly, neither recall nor clustering differed significantly across the three procedures, indicating that EFR instructions did not appreciably affect item availability or contextually based search.

Examinations of search accuracy revealed that participants could selectively retrieve items from a target list at above chance level. At the beginning of the recall period, search accuracy was typically extremely high indicating successful reinstatement of the target context. As the recall period progressed accuracy steadily declined. Interestingly, there was no difference in constrained search between recall of List 1 and List 2, either collapsed across all output positions or at different stages of the recall period. This suggests that participants are equally adept at reinstating past contexts as they are the present context. Furthermore, participants who scored higher for constrained search accuracy also exhibited greater clustering in item generation indicating an important role of context in searching for targets.

One issue with List membership is that it is theoretically possible to retrieve targets without reinstating source context. A participant could simply use incidentally encoded inter-item temporal associations to isolate all items in the trial and retrieve the correct list without actually excluding the incorrect one, at least in the case of List 2. To address this issue, two versions of each experiment were run: One with visual presentation of items, the other with auditory. The literature shows us that the Auditory modality appears to benefit from a better representation of temporal information than the Visual modality (Glenberg & Swanson, 1986). Therefore, if constrained search in List membership experiments is largely driven by inter-item temporal associations or rehearsal strategies, then constrained search should be superior for the Auditory modality than Visual. No significant effect of Modality on constrained search was found. However, this lack of an effect of Modality may be as a result of the task. Modality effects in free recall tend to manifest as enhanced recency effects for the Auditory modality (Murdock & Walker, 1969), which tend to be extinguished in delayed-free recall experiments (Postman & Phillips, 1965) such as

those presented in this thesis. Therefore, ultimately participants may not have derived any benefit in terms of constrained search from the Auditory modality.

Target monitoring was near ceiling for most of the recall period, with a small drop in accuracy in the last few output positions. Participants could also monitor source intrusions at above chance level. Interestingly, source intrusion monitoring accuracy was below chance at output position 1, rose sharply to above chance at output position 2, and rose again to ceiling near the end of the recall period.

The source monitoring framework (Johnson et al. 1993) states that participants have access to different types of source information at different times. Perceptual information regarding a stimulus such as its visual or auditory features can be available automatically as the memory is retrieved. Other source judgments which require conscious thought, for example judging the feasibility that an item came from List 1 or List 2 take longer and may require supporting memories. It could be argued that there are no distinguishing perceptual differences between List 1 and List 2; therefore, all source judgments have to be made using these slower conscious processes. As recall begins rapidly and slows exponentially throughout the recall period, (Wixted & Rohrer, 1993) it is likely that at the first output position participants may have no source information available to make a judgment, and just assume that they have retrieved a target. As recall slows, participants have more time to make conscious source monitoring judgments and performance improves.

Chapter 3 aimed to expand on the initial EFR study conducted in Chapter 2, to investigate factors which may influence the effectiveness of constrained search and monitoring, in Mixed-lists of different sources. In Mixed-lists, inter-item associations are unhelpful in retrieving targets; therefore, the participant must attempt to reinstate source context to focus search. This gave a clearer picture of the effectiveness of

constrained search, and the role of source given that confounds related to temporal context were now
removed.

The first study examined the effect of Source Similarity on constrained search and monitoring. The Source Monitoring Framework (Johnson et al. 1993) describes how monitoring judgments are less accurate when the sources to be discriminated are more similar than when they are less similar. Constrained search should also be poorer when sources are more similar, as the retrieval cue for targets will be a closer contextual match to the incorrect source, than when two sources are less similar (Polyn et al. 2009a). Indeed the clustering literature suggests this. Hintzman et al. (1972) found stronger source clustering in Mixed-lists of items presented in different modalities (auditory and visual), than two sources within the same modality (male and female voices).

In the present thesis, an EFR study was run using the same Similarity manipulation as Hintzman et al. (1972). High-similarity trials were Mixed-lists of items presented in either a male or female voice, and Low-similarity trials comprised items presented in either the visual or auditory modality. Surprisingly there was found to be no effect of Similarity on constrained search, measured by both aggregate data and retrieval dynamics. It is possible that the participants may have noticed additional distinguishing features between the two sources in High-similarity trials that were not experimentally controlled such as regional accent. Alternatively, participants may have developed their own rehearsal strategy unknown to the experimenter which assisted in distinguishing the sources. Either of these methods could be used to develop more distinctive search cues, and assist in reinstating the target context, improving constrained search accuracy.

However, as expected monitoring of source intrusions was more accurate in Low similarity lists than High-similarity trials, demonstrating that the Similarity manipulation was effective. As previously stated, participants have limited time to make monitoring decisions early in the recall. Monitoring judgments requiring additional cognitive operations such as “What accent did the voice have?” are slower; therefore, participants may not have had time to use these operations to assist with monitoring judgments early in the recall period. Monitoring judgments would be solely made based on perceptual information pertaining to the item, in this instance voice gender. This is supported by the output dynamics data for intrusion monitoring in Experiment 3.1, where there is better performance for Low-similarity trials early on in the recall period. However at middle output positions, performance of the two Similarity conditions equalises. One could also argue that some participants may have completely neglected source early in the recall period, where they assume that rapid retrievals must be targets. In essence, no source monitoring has occurred.

Following on from Experiment 3.1, the concept of Context Dependency was explored to investigate if participants can utilise additional task-irrelevant source information to assist with constraining search and monitoring. Initially, a pilot study was run to determine which sources participants could successfully search for and monitor. Experiment 3.2 demonstrated that participants can search for targets at above chance level using Font Colour, Font Size and Background picture as retrieval cues. Retrieval dynamics for search showed the same pattern as List membership and Modality/voice gender, with high search accuracy at the beginning of the recall period, declining with output position. Monitoring of these source features in addition to screen location was also well above chance. Source intrusion monitoring dynamics

showed the same pattern as List membership and Modality/voice gender. Accuracy was very poor early on, rising to ceiling at the end of the recall period.

To investigate Context Dependency, participants studied lists of items presented in four different contexts simultaneously: One of two Font Colours, Font Sizes, Screen Locations and Background pictures. At the end of each list participants were asked to retrieve targets based on one of these four context, for example “recall all the items printed in red”. Context Dependency was manipulated as the probability of any given source predicting the sources from the other three contexts. In a Dependent condition, sources were totally predictive of one another. For instance, all red items were presented at the top of the screen, in large font and against the background of a tiger. In the Independent condition, each item was randomly assigned a source from all four contexts, meaning that no source predicted another, yielding a Dependency probability of 0.5. In a Partially Dependent condition, the probability of Dependency was 0.75.

It was found that participants were not able to successfully constrain search in any of the Dependency conditions. Retrieval dynamics support this by demonstrating that search accuracy fluctuates around chance throughout the recall period. Target monitoring was significantly above chance for all Dependency conditions. However participants were unable to monitor source intrusions at above chance level in the Dependent or Partially dependent conditions. Further, evidence for above chance source intrusion monitoring in the Independent condition was weak. Source intrusion monitoring dynamics show an increase to above chance only at the end of the recall period. This could be source neglect, as participants believe that they will likely retrieve source intrusions, and; therefore, attribute these items to the incorrect source without engaging in monitoring.

In reality it would be presumptuous to state that Context Dependency truly has no effect on ability to constrain search. Experiment 3.2 demonstrated that these same source features can be used as cues to constrain search in isolation; however, when the participant is required to simultaneously study four sources for each item they can no longer constrain search. The lack of ability to monitor source intrusions irrespective of Dependency condition seems to indicate that source was poorly encoded. This suggests that the task was simply too challenging, and would benefit from the removal of either one or two contexts to gain a more clear picture of the effect of Context Dependency.

Chapter 3 also lends support for EFR as a viable and reliable way of measuring constrained search. As previously stated, one of the issues surrounding EFR is confounds of selective reporting. One way to test the impact of this is to investigate whether EFR can successfully detect predictable differences between contexts. One strong prediction is that constrained search accuracy should be significantly worse in Mixed-lists than for List membership. If selective reporting is a significant confound, then there should be no difference between these two types of context. Comparisons between List membership in Experiment 2.3 and Mixed-lists in Experiment 3.1 tentatively suggest poorer constrained search for Mixed-list contexts than List membership, demonstrating that at least in this instance, selective reporting did not obscure the effect. However, this is not to say that selective reporting did not mask more nuanced effects on constrained search, such as the predicted difference between High and Low-similarity trials which was not supported by the data. Chapter 4 aimed to further investigate this issue.

Chapter 4 applied a curve fitting methodology to assess constrained search accuracy, which aims to estimate the size of a participant's search set by fitting recall

latencies. Making the assumption that items are sampled at a constant rate from a search set with replacement, recall latencies should progressively slow with the number of items output. An exponentially-modified-Gaussian curve (ex-Gaussian) was used, as this has shown to be a good fit to recall latencies in previous studies (Rohrer & Wixted, 1994; Wixted & Rohrer, 1993). This is a convolution of the Gaussian and exponential distributions. The distribution comprises three parameters. The most important to this investigation was the exponential rate parameter tau which indexes search set size. The other parameters are mu and sigma which relate to the onset of recall and variability in the onset of recall respectively. Successful constrained search using the latency methodology is indicated by a significantly smaller estimate of tau for recall of a single source than recall of both sources in a trial. Note that there is no instruction with this paradigm to report source intrusions; therefore, the methodology is immune to the selective reporting issues suspected to be present in EFR.

Prior to discussion of these studies it is important to note that although the data did take the basic form of the ex-Gaussian, none of the ex-Gaussian fits were a good mathematical fit to the data as assessed by chi-squared goodness of fit tests. Therefore, the obtained parameter estimates should be treated with great caution. Behavioural data from these experiments were also analysed as a secondary assessment of the feasibility of the parameter estimates. If the parameter estimates are inconsistent with the underlying recall data from the same experiment, then the conclusions drawn from parameter estimation are not representative of participant behaviour.

The first experiment of Chapter 4 was an independent replication of Experiment 2.3 (List membership). On each trial, participants studied two lists of items, and were then asked to verbally recall List 1, List 2 or Both lists. Tau was significantly

smaller for recall of List 1 or List 2 than Both lists, indicating successful constrained search in both cases. In fact, the value of tau for List 2 was roughly half of that for Both lists. Crucially this was consistent with the underlying recall data, so these estimates are feasible. Interestingly, tau was significantly smaller for List 2 than List 1, indicating that search set size was larger for List 1 than List 2. The retrieval cue for List 1 appears to be somewhat more noisy than the retrieval cue for List 2. This is particularly interesting as it is an effect which was not detected by EFR. Experiment 2.3 found no difference in constrained search between Lists 1 and 2. This potentially indicates a lack of sensitivity in EFR to detect more subtle effects on constrained search, possibly due to selective reporting.

Experiment 4.2 was an independent replication of Experiment 3.2. This experiment aimed to observe whether the recall latency methodology replicated the finding that participants can still constrain search in Mixed-lists. An additional prediction was that constrained search should be poorer for Mixed-lists than for List membership. For this experiment, participants studied a list of items, each presented in one of two screen locations. They were then asked to recall only the items presented in one of the two screen locations, or all of the items. Tau was significantly smaller for recall of a single source than both sources indicating successful constrained search in Mixed-lists.

When Mixed-list context was compared with List membership, tau was smaller for List membership than for Mixed-list context when a single source was recalled. However when both sources were recalled tau was significantly larger for List membership than Mixed-lists. Importantly, the difference in tau between a single source and both sources was much smaller for Mixed-lists than for List membership indicating that participants were worse at constraining search in Mixed-lists,

supporting predictions. Interestingly the main origin of this effect appears to be the number of items searched when both sources are recalled, as the difference between the two contexts was much larger when both sources were being recalled than a single source. This was supported by the behavioural data, implying that Mixed-list sources are a weaker retrieval cue than temporal context for List membership.

The final experiment of Chapter 4 was an independent replication of Experiment 3.1. The aim was to examine any potential effect of Source Similarity on ability to constrain search. The equivalent EFR experiment found no effect of Similarity at all. However, it may be the case that EFR is not sensitive enough to detect such nuanced effects due to selective reporting. Latency measures were used in an attempt to clarify this point. Participants studied lists if items presented in one of two sources. In the High-similarity trials, items were either spoken in a male voice or a female voice. In Low-similarity trials, items were either presented on the computer screen or through speakers.

The results from this experiment are largely difficult to explain. For instance there is no reason why participants should search more items when recalling a single source than both sources in Low-similarity trials, whereas the opposite was true for High-similarity trials. It is also difficult to explain why, when recalling a single source, search set size was larger for Low-similarity trials than High-similarity trials. However, the behavioural data from this experiment is largely consistent with the results from EFR, indicating that the parameter estimates from the ex-Gaussian fits do not represent true behaviour.

Given the relative success of Experiment 4.1 in replicating the findings of Experiment 2.3, and perhaps even revealing effects not detected by EFR, this shows promise for latency analysis as a complementary method for measuring search

accuracy control, along with EFR. The poor ex-Gaussian fits may be attributable at least in part to the way the data were analysed. Recall for individual participants was too poor to perform analysis at the single subjects level. Therefore, participant data from individual conditions were pooled as if they originated from a single person before the ex-Gaussian was fit. Essentially this means that none of the ex-Gaussian fits are representative of any given participant. When this was examined, data aggregation was found to affect different fits to different degrees. Noise in the underlying data would appear to play a large part in the ex-Gaussian fits being poor. To combat the data aggregation issue, future work in this area should run experiments over multiple sessions, so that participants complete multiple trials of the same conditions. This would hopefully provide more data, sufficient to conduct individual subjects analysis.

The final exercise of this thesis looked to examine the EFR search dynamics data in more detail. Specifically, to see whether it was possible to model the drop in search performance with output position simply as a function of a falling base rate of targets over the recall period. Retrieval was modelled as a sampling with replacement process similar to that described by Wixted and Rohrer (1993). Search set size is determined by a parameter n and comprises an equal number of targets and wrong-source items. Items are sampled with replacement from the search set with a static probability of retrieving a target throughout the recall period, determined by a free parameter p. Items continue to be sampled in this way until a certain number of items which have already been sampled are sampled again consecutively (not necessarily the same item). The number of consecutive repeats allowed before retrieval terminates is determined by the free parameter s.

In the first, most simple iteration of the model, no repetitions appeared in the generated output; they were simply replaced in the search set and monitored as a

repetition. This is feasible, as participants were not given any specific instructions for when a previously retrieved item was generated. Even though they were told to write all correct and incorrect information that came to mind, they may still have assumed that repetitions should not be written again. Monitoring of repetitions in this first iteration was also deemed to be perfect, while source monitoring played no role whatsoever.

To examine the model's ability to describe the recall dynamics data, the model was fitted to the search dynamics from Experiment 2.3 collapsed across Modality and List membership. For the model to be a good fit to the data, it was necessary that the model adequately predicted the pattern of search accuracy across the recall period, recall metrics such as the mean number of targets and source intrusions generated, and the proportion of total data contributing to each output position (dropout rate).

Model iteration 1, with perfect repetition monitoring and no recorded repetitions was fairly successful at predicting the pattern of decline in search accuracy for the List membership data, and estimations of target and source intrusions were reasonably accurate albeit slightly underestimated. However the predicted dropout rate was far steeper than the observed dropout rate.

A second model iteration was developed to address issues of underestimating dropout rate. A new free parameter 'repm' was introduced to control the accuracy of repetition monitoring. This allowed for monitoring errors whereby novel items could be monitored as repetitions and vice versa. This modification had no appreciable effect on the fit to the dynamics data. Although this model iteration did provide a better fit to recall metrics and the dropout rate. Unfortunately this could not be attributed to 'repm', as the best fitting parameter estimate for 'repm' was 1, or perfect repetition monitoring. Instead, the improved fit to the recall metrics and dropout rate must be

due to a more liberal stopping rule, controlled by parameter ‘s’. This has ramifications for the model’s reliability. Given that the best fitting parameter estimate for ‘repm’ was 1, both model iterations had essentially the same assumptions, but yielded different best fitting parameter estimates. This is likely due to an insufficient sample size for the simulations, which is discussed further in Chapter 5. However, despite this it is still possible to conclude that a simple explanation based on a reducing pool of targets alone is unlikely to account for the patterns of search dynamics data observed in Experiment 2.3.

In general it would seem that constrained search is at least in part accomplished by reinstatement of the target context at retrieval. Models such as CMR2 (Lohnas et al. 2015) which do not describe such a mechanism cannot explain for instance why there is no difference in constrained search accuracy between recall of the most recent list and a previous list. The role of context is also demonstrated by significant differences in constrained search accuracy between different types of context. Constrained search was significantly greater in List membership than Mixed-lists, owing to the fact that the two sources are also separated by time, whereas this is not the case for Mixed-lists. However, it is also evident that there are other factors at play when constraining search aside from context reinstatement. The evidence for this is that constrained search was largely resistant to manipulations within a single experiment, such as Source similarity. The participant may be using other internally generated retrieval cues which cannot be experimentally accounted for, that can help in distinguishing between sources, as detailed by Meta-RAR (Goldsmith, 2016).

Source monitoring performance appeared to be quite predictable and largely followed the principles of the Source Monitoring Framework (Johnson et al. 1993).

Participants appear to neglect source at the first output position when retrieval is rapid, and engage more in source monitoring when more source information is available to them as retrieval slows. Monitoring also appears to respond to other principles detailed by the framework, namely Source Similarity.

Future research in this area could explore other aspects of memory not covered in this thesis. For instance how does source strength/distinctiveness influence constrained search? The von Restorff isolation effect is a well-known phenomenon in the memory literature whereby individual items that are distinctive on some dimension, have a better probability of recall than items which are the same on that dimension (Karis et al., 1984). For example in a list comprising one item printed in red font and ten items printed in blue font, the red item has a greater probability of recall than any of the blue items, presumably due to the increased distinctiveness and strength of the 'red' source. If this isolation paradigm were applied to EFR, one would expect that constrained search would be more accurate when recalling the source with fewer items than the source with more items. Due to the greater relative strength of the more distinctive source, it is more likely to intrude when it is task irrelevant than the less distinctive source.

Additionally, one should look to optimise the design for the latency analyses conducted in Chapter 4, as this shows great promise as a complementary method to study constrained search. Finally, as an extension to the modelling work conducted in Chapter 5, one could explore a potential role of source monitoring in describing the retrieval dynamics, by simulating situations where the stopping rule is determined by source intrusions, or by simulating selective reporting, whereby participants deliberately withhold source intrusions that they know are incorrect.

References

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

Bjork, R.A., & Whitten, W.B. (1974).Recency-sensitive retrieval processes in long term recall. *Cognitive Psychology, 6,* 173-189.

Bousfield, W.A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology, 49,* 229-240.

Bousfield, W.A., Cohen, B.H., & Silva, J.G. (1956). The extension of Marbe's law to the recall of stimulus-words. *American Journal of Psychology, 69,* 429-433.

Bousfield, W.A., & Rosner, S.R. (1970). Free vs uninhibited recall. *Psychonomic Science, 20,* 75-76.

Bousfield, W.A., & Sedgewick, C.H.W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30,* 149-165.

Bousfield, W.A., Sedgewick, C.H.W., & Cohen, B.H. (1954). Certain temporal characteristics of the recall of verbal associates.

Bousfield, W.A., Whitmarsh, G.A., & Esterson, J. (1958). Serial position effects and the Marbe effect in the free recall of meaningful words. *Journal of General Psychology, 58,* 255-262.

Brown, A.S. & Murphy, D.R. (1989). Cryptomnesia: Delineating inadvertent plagiarism. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15,* 432-442.

Brown, S.C., Conover, J.N., Flores, L.M., & Goodman, K.M. (1991). Clustering and recall: Do high clusterers recall more than low clusterers because of clustering? *Journal of Experimental Psychology: Learning, Memory and Cognition, 17,* 710-721.

Cofer, C.N., Bruce, D.R., & Reicher, G.M. (1966). Clustering in free recall as a function of certain methodological variations. *Journal of Experimental Psychology, 71,* 858-866.

Del Missier, F., Sassano, A., Coni, V., Salomonsson, M., & Mäntylä, T. (2018). Blocked vs. interleaved presentation and proactive interference in episodic memory. *Memory, 26,* 697-711.

Doerksen, S., & Shimamura, A.P. (2001). Source memory enhancement for emotional words. *Emotion, 1,* 5-11.

Ecker, U.K.H., Zimmer, H.D., & Groh-Bordin, C. (2007). The influence of object and background colour manipulations on the electrophysiological indices of recognition memory. *Brain Research, 1185,* 221-230.

Einstein, G.O., & McDaniel, M.A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16,* 717-726.

Frost, N. (1971). Clustering by visual shape in the free recall of pictorial stimuli. *Journal of Experimental Psychology, 88,* 409-413.

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

Glenberg, A.M. (1987). Temporal context and recency. In D.S. Gorfein, R.R. Hoffman (Eds.), *Memory and Learning: The Ebbinghaus Centennial Conference* (pp. 173-190). Hillsdale NJ, US: Lawrence Erlbaum Associates, Inc.

Glenberg, A.M., Bradley, M.M., Kraus, T.A., & Renzaglia, G.J. (1983). Studies of the long-term recency effect: Support for a contextually guided retrieval hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9,* 231-255.

Glenberg, A.M., & Swanson, N.G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12,* 3-15.

Goldsmith, M. (2016). Metacognitive quality-control processes in memory retrieval and reporting. In J. Dunlosky, S. Tauber (Eds.), *The Oxford Handbook of Metamemory* (pp. 357-384). Oxford, England: Oxford University Press.

Harbison, J.I., Davelaar, E.J., Yu, E.C., Hussey, E.K., & Dougherty, M.R. (2013). Intrusions and the decision to terminate memory search. In M. Knauff., M. Pauen., N. Sebanz., I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 549–554). Austin, TX: Cognitive Science Society.

Heathcote, A., Popiel, S.J., & Mewhort, D.J.K. (1991). Analysis of response time distributions: An example using the stroop task. *Psychological Bulletin, 109,* 340-347.

Hintzman, D.L. (2016). Is memory organized by temporal contiguity? Memory & Cognition, 44, 365-375.

Hintzman, D.L., Block, R.A., & Inskeep, N.R. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behaviour, 11,* 741-749.

Hintzman, D.L., Summers, J.J., & Block, R.A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 31-40.

Hollins, T.J., Lange, N., Berry, C.J., & Dennis, I. (2016). Giving and stealing ideas in memory: Source errors in recall are influenced by both early-selection and late-correction retrieval processes. *Journal of Memory and Language, 88,* 87-103.

Hollins, T.J., Lange, N., Dennis, I. & Longmore, C.A. (2016). Social influences on unconscious plagiarism and anti-plagiarism. *Memory, 24,* 884-902.

Howard, M.W., & Kahana, M.J. (1999).Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25,* 923-941.

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

Hudson, R.L. (1968). Category clustering as a function of level of information and number of stimulus presentations.*Journal of Verbal Learning and Verbal Behaviour, 7,* 1106-1108.

Hudson, R.L. & Dunn, J.E. (1968). A major modification of the Bousfield (1966) measure of category clustering. *Behavior Research Methods and Instrumentation, 1,* 110-111.

Jang, Y. & Huber, D.E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34,* 112-127.

Johnson, M., Hashtroudi, S., & Lindsay. D.S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3-28.

Kahana, M.J. (1996). Associative retrieval processes in free recall. *Memory and Cognition, 24,* 103-109.

Kahana, M.J. (2017). Memory search. In J.H. Byrne (Ed.), *Learning and Memory: A comprehensive Reference* (pp 181-200). Oxford: Academic Press.

Kahana, M.J., Dolan, E.D., Sauder, C.L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *The Journals of Gerontology: Series B, 60,* 92-97.

Kahana, M. J., Howard, M. W., & Polyn, S. M. (2008). Associative Retrieval Processes in Episodic Memory.*Psychology*.3. https://surface.syr.edu/psy/3.

Karis, D., Fabiani, M., & Donchin, E. (1984). ”P300” and memory: Individual differences in the von Restorrf effect. *Cognitive Psychology, 16,* 177-216.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517,

Koriat, A., Goldsmith, M., & Halamish, V. (2008). Controlled processes in voluntary remembering. In J. Byrne & H.L. Roediger III (Eds.), *Learning and Memory: A Comprehensive Reference: Vol. 2. Cognitive Psychology of Memory* (pp. 307-324). Oxford, England: Elsevier.

Laming, D. (2009). Failure to recall. *Psychological Review, 116,* 157-186.

Landau, J.D., & Marsh, R.L. (1997). Monitoring source in an unconscious plagiarism paradigm. *Psychonomic Bulletin and Review, 4,* 265-270

Lohnas, L.J., & Kahana, M.J. (2014). Compound cueing in free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 40,* 12-24.

Lohnas, L.J., Polyn, S.M., & Kahana, M.J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review, 122,* 337-363.

Maylor, E.A., Chater, N., & Jones, G.V. (2001). Searching for two things at once: Evidence of exclusivity in semantic and autobiographical memory retrieval. *Memory and Cognition, 29,* 1185-1195.

Miller, J.F., Lazarus, E.M., Polyn, S.M., & Kahana, M.J. (2013). Spatial clustering during memory search. *Journal of Experimental Psychology: Learning, Memory and Cognition, 39,* 773-781.

Miller, J.F., Weidemann, C.T., & Kahana, M.J. (2012). Recall termination in free recall. *Memory and Cognition, 40,* 540-550.

Mintzer, M.Z., & Snodgrass, J.G. (1999). The picture superiority effect: Support for the distinctiveness model. *The American Journal of Psychology, 112,* 113-146.

Morey, R.D., & Rouder, J.N. (2018). BayesFactor: Computation of Bayes Factors for common designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Mulligan, N.W. (2004). Generation and memory for contextual detail. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30,* 838-855.

Murdock, B.B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482-488.

Murdock, B.B., & Walker, K.D. (1969). Modality effects in free recall. *Journal of Verbal Learning and Verbal Behaviour, 8,* 665-676.

Neath, I., & Crowder, R.G. (1990). Schedules of presentation and temporal distinctiveness in human memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16,* 316-327.

Nelder, J.A., & Mead, R. (1965). A simplex method for functional minimization. *Computer Journal, 7,* 308-313.

Nilsson, L.G. (1974). Further evidence for organization by modality in immediate free recall. *Journal of Experimental Psychology, 103,* 948-957.

Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology, 5,* 176-206.

Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Hochenberger, R., Sogo, H., Kastman, E., & Lindelof, J.K. (2019). PsychoPy2: Experiments in behaviour made easy. *Behavior Research Methods, 51,* 195-203.

Pillemer, D.B., Goldsmith, L.R., Panter, A.T., & White, S.H. (1988). Very long-term memories of the first year in college. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14,* 709-715.

Polyn, S.M., Norman, K.A., & Kahana, M.J. (2009a). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116,* 129-156.

Polyn, S.M., Norman, K.A., & Kahana, M.J. (2009b). Task context and organisation in free recall. *Neuropsychologia, 47,* 2158-2163.

Postman, L., & Phillips, L.W. (1965).Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology, 17,* 132-138.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1981). Search of associative memory. *Psychological Review, 88,* 93-134.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

Roenker, D.L., Thompson, C.P., & Brown, S.C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76,* 45-48.

Rohrer, D. & Wixted, J.T. (1994). An analysis of latency and interresponse time in free recall. *Memory and Cognition, 22,* 511-524.

Santa, J.L., Ruskin, A.B., Snuttjer, D., & Baker, L. (1975). Retrieval in cued recall. *Memory and Cognition, 3,* 341-348.

Sederberg, P.B., Miller, J.F., Howard, M.W., &Kahana, M.J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition, 38,* 689-699.

Shiffrin, R.M. (1970). Forgetting: Trace erosion or retrieval failure? *Science, 168,* 1601-1603.

Smith, S.M. (1982). Enhancement of recall using multiple environmental contexts during learning. *Memory and Cognition, 10,* 405-412.

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

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174-215.

Starns, J.J., & Hicks, J.L. (2005). Source dimensions are retrieved independently in multidimensional monitoring tasks. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31,* 1213-1220.

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology, 18,* 813-847.

Steyvers, M., Shiffrin, R.M., & Nelson, D.L. (2004).Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Cognitive psychology and its applications:Festschrift in honor of Lyle Bourne,*

*Walter Kintsch, and Thomas Landauer.*Washington, DC: American Psychological Association.

Tzeng, O.J.L. (1973). Positive recency effect in a delayed free recall. *Journal of Verbal Learning and Verbal Behaviour, 12,* 436-439.

Unsworth, N., Brewer, G.A., & Spillers, G.J. (2010). Understanding the dynamics of correct and error responses in free recall: Evidence from externalized free recall. *Memory & Cognition, 38,* 419-430.

Unsworth.N., Brewer, G.A., & Spillers, G.J. (2013). Focusing the search: Proactive and retroactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 39,* 1742-1756.

Unsworth, N., Spillers, G.J. & Brewer, G.A. (2012). Evidence for noisy contextual search: Examining the dynamics of list before last recall. *Memory, 20,* 1-13.

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.J., Derks, K., Dablander, F., Gronau, Q.F., Kucharsky, S., Gupta, A.R.K.N., Sarafoglu, A., Voelkel, J.G., Stefan, A., Ly, A., Hinne, M., Matzke, & Wagenmakers, E.J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *Annee Psychologique, 120,* 73-96.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*, 289–335.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted (Vol. Ed.) & H. Pashler (Series Ed.). *Stevens' Handbook of Experimental Psychology (3rd*

Edition), Volume 4: Methodology in Experimental Psychology (pp. 461-516). New York: Wiley Press.

Watkins, M.J., & Gardiner, J.M. (1979). An appreciation of generate-recognize theory of recall. *Journal of Verbal Learning and Verbal Behaviour, 18,* 687-704.

Wickens, D.D. (1973). Some characteristics of word encoding. *Memory and Cognition, 1,* 485-490.

Wixted, J.T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19,* 1024-1039.

Appendix A:

A1.1. *Bayesian post hoc analysis*

All Bayesian post-hoc analyses in this thesis were conducted using a method proposed by van den Bergh et al. (2020). This approach relates the probability that all means are equal under the null hypothesis (H_0), to the probability of any two means being equal. The prior probability can then be adjusted according to the number of conditions in the design to correct for multiplicity.

The premise is that any condition mean is equal to the grand mean μ with a probability τ, or different from the grand mean with a probability $1-\tau$. The probability that any two given means are equal is $p(\mu_1 = \mu_2) = p(\mu_1 = \mu) \times p(\mu_2 = \mu) = \tau^2$. Therefore, the probability that all y means are equal under the null hypothesis H_0 is expressed as: $p(\mu_1 = \mu_2 \dots \mu_x) = p(\mu_1 = \mu) \times p(\mu_2 = \mu) \times \dots \times p(\mu_y = \mu) = \tau^y$. Solving for τ gives $\tau = p(H_0)^{1/y}$. Now the prior probability that any two conditions are equal can be represented in terms of the prior probability that all means are equal under H_0. The equation for this is: $p(\mu_i = \mu_j) = \tau^2 = p(H_0)^{2/y}$. As an example if there are four conditions in an experimental design, the probability that all condition means will be equal is $p(H_0)^{2/4} = 0.5$. Therefore the probability that any pair of condition means will be equal is $\sqrt{0.5}$. The prior odds are then calculated as $(1-\sqrt{0.5}) / \sqrt{0.5} = 0.41$. Once the prior odds have been calculated, a Bayesian t-test is run for each comparison, and the resulting Bayes Factor is multiplied by the prior odds to gain the posterior odds. It is these posterior odds that are reported and interpreted for all post-hoc analyses.

Appendix B

B1.1. *Thesis equations*

Equations are numbered such that the first number corresponds to the chapter where the formula is first applied, and the second number is the number of the equation within the chapter. For instance Equation 2.1 first appears in Chapter 2 and is the first equation presented in Chapter 2.

2.1 - Proportion of items correctly recalled per list for experiments 2.1 and 2.2.

$$PcRecall = \frac{n_l}{10}$$

Where n_l is the number of items recalled by a participant in list number l.

2.2 - Adjusted ratio for clustering (clustering score per trial for experiments 2.1, 2.2 and 2.3).

$$ARC = \frac{R - E(R)}{maxR - E(R)}$$

Where R represents the number of observed repetitions.

2.3 - Maximum number of same category repetitions for a recall output (denominator term

of ARC).

$$maxR = N - k$$

Where N is the total number of items recalled by the participant, and k is the number of sources in the trial (2 in all cases).

2.4 - Expected number of repetitions given chance clustering (numerator and denominator

term of ARC).

$$E(R) = \frac{\sum_i n_i^2}{N} - 1$$

Where n_i is the number of items recalled from category (source) i and N is as before.

2.5 - Source monitoring accuracy per trial for Experiment 2.2.

$$PcMonitor = \frac{A_a + B_b}{A_a + A_b + B_a + B_b}$$

Where A and B represents the list that the item was presented in, and a and b is the participant's monitoring response.

2.6 - Proportion of items correctly recalled per trial for Experiments 2.1 and 2.2.

$$PcRecall = \frac{N}{20}$$

Where N is the total number of items recalled in the trial.

2.7 - Proportion of correct source items recalled in Experiment 2.3

$$PcRecall = \frac{T}{10}$$

Where T is the total number of targets generated.

2.8 - Proportion of total items correctly recalled that were from the target source per trial (overall search accuracy). Used throughout the thesis.

$$PcSource = \frac{T}{T + S}$$

Where T and S represent the number of targets and source intrusions (SI) generated respectively.

2.9 - Target monitoring accuracy per trial. Used throughout thesis

$$Target\ monitoring = \frac{T_t}{T_t + T_s}$$

2.10 - Source intrusion monitoring accuracy per trial. Used throughout thesis.

$$Source\ intrusion\ monitoring = \frac{S_s}{S_s + S_t}$$

For equations 2.9 and 2.10, T and S are targets and source intrusions, and t and s are the participant's monitoring response. T_t is a target correctly monitored as a target, T_s is a target incorrectly monitored as a source intrusion, S_s is a source intrusion correctly monitored as a source intrusion, and S_t is a source intrusion incorrectly monitored as a target.

3.1 - Proportion of interference intrusions per second condition in trials 3 and 4 of Experiment 3.1 and Experiment 4.3.

$$P_{int} = \frac{Int_3 + Int_4}{n_3 + n_4}$$

Int_3 and Int_4 are male voice interfering items in the trial 3 and 4 recall outputs respectively, and n_3 and n_4 are the total number of generated items in trial 3 and trial 4 respectively. The same formula is used for Experiment 4.3; however, the recall output only contains overtly recalled items rather than all items generated.

3.2 - Proportion of targets generated in Experiments 3.2 and 3.3 per trial.

$$PTarget = \frac{t}{T}$$

Where t is the number of targets generated by the participant and T is the total number of targets in the trial.

3.3 – Proportion of source intrusions generated in Experiments 3.2 and 3.3 per trial.

$$PSI = \frac{s}{S}$$

Where s is the number of Source intrusions generated and S is the total number of presented wrong source items in the trial.

4.1 - Exponentially-modified Gaussian (ex-Gaussian) probability density function for analysis of recall latencies.

$$f(t) = \frac{e^{-\frac{t-\mu}{\tau}+\sigma^2/2\tau^2}}{\tau\sqrt{2\pi}} \int_{-\infty}^{\frac{t-\mu}{\sigma}-\sigma/\tau} e^{-\frac{y^2}{dy}}$$

Where τ represents the average time of the retrieval phase and μ and σ are the mean and standard deviation of duration of the search set establishment phase. For latency analysis mu and sigma are established by Gaussian kernel estimation, and tau by Maximum likelihood estimation.

4.2 - Formula to ensure that only plausible parameter values are explored for Gaussian kernel estimation.

$$\sqrt{\frac{\min(x-M)^2}{n-1}} \leq \sigma \leq S$$

Where M and S are the mean and standard deviation of the data.

4.3 - Joint likelihood of a set of parameter values given data (Maximum likelihood estimation).

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{k=1}^{k} L(\boldsymbol{\theta}|y_k)$$

Where Θ is a set of parameter values, in this case values of μ, σ and τ, y is a sample of data and k represents individual data points.

4.4 - Negative sum of log likelihood for Maximum likelihood estimation

$$-\ln L(\boldsymbol{\theta}|\boldsymbol{y}) = -\sum_{k=1}^{K} \ln L(\boldsymbol{\theta}|\boldsymbol{y})$$

All variables are the same as Equation 4.3.

4.5 - Proportion of targets overtly recalled in Experiment 3.2 (EFR) per trial.

$$PTarget = \frac{t_t}{T}$$

Where t_t is the number of targets generated that were written in the 'target' box, and T is the total number of targets in the trial.

4.6 - Proportion of source intrusions overtly recalled in Experiment 3.2 (EFR) per trial.

$$PSI = \frac{s_t}{S}$$

Where s_t is the number of source intrusions generated that were written in the 'target' box, and S is the total number of wrong-source items in the trial.

4.7 - Proportion of items overtly recalled in Experiment 3.2 (EFR) per trial.

$$PRecall = \frac{t_t + s_t}{N}$$

Where t_t and s_t are the same as Equations 4.5 and 4.6, and N is the total number of items in the trial.