Metacognitive awareness of associative learning: What underlies delayed judgments-of-learning?

Radka Jersakova

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Psychology

October 2016

INTELLECTUAL PROPERTY AND PUBLICATIONS

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

ACKNOWLEDGEMENTS

I have been extremely lucky in the course of my PhD to have worked with a number of fantastic supervisors and mentors who together made this a truly invaluable experience. First and foremost I want to thank Richard Allen as without his guidance, insight and tremendous patience it is hard to imagine this thesis being completed. Secondly, I want to thank Céline Souchay for her enthusiasm for research and her unwavering trust in my abilities, which always inspired me to want to achieve more. Akira O'Connor and Chris Moulin, while never my official supervisors, always supported me, ready to answer any questions I might have and gave me opportunities to develop ideas beyond this thesis, for which I am incredibly grateful. I also want to thank Amanda Waterman and Charity Brown for general guidance and encouragement and Janet Metcalfe for welcoming me in her lab in New York.

I could not have done any of this thesis without the support of the many wonderful colleagues I have met along the way, especially here in Leeds and in Dijon. Special thanks goes to my various office mates, particularly Callum, Oscar, Bex, Emma, Helen, Katy and Rose who patiently listened to me talk through ideas and complain about problems while later admitting they did not often know what I was talking about and to Christophe, Laura and Mathieu for making me feel at home in a country where, to begin with, I barely spoke the language.

I also want to thank all my friends scattered around the globe for their positive energy and for helping me remember there is life outside of academia. Last but not least, I want to thank my family who always supported me in all my endeavours and who, especially in these last few months of writing up, looked after me with unimaginable patience and care.

ABSTRACT

Cognitive processes, such as memory, are accompanied by metacognitive states of awareness that allow for evaluation of their function. Across seven experiments we employed the delayed judgment-of-learning (JOL) paradigm with healthy young adults to examine metacognitive monitoring of learning. After studying cue-target word-pairs, participants were presented with the studied cues and predicted their ability to retrieve the target on a subsequent memory test. The key question of interest was the nature of the underlying processes guiding such judgments with a focus on how they relate to memory. The delayed JOL literature has assumed that it is an absolute judgment, based on the ease of access to the target item. Chapters 2 and 3 manipulated target- and cue-related variables and investigated their influence on memory and metamemory. The results showed delayed JOLs are also sensitive to memory for contextual information about the target (Chapter 2) and the level of familiarity with the cue term (Chapter 3). This is strengthened by results from Chapter 4 in which participants provided written justifications of their JOL responses without any experimental manipulations of the learned material. Analysis of these responses confirmed that both cue- and target-related information influences delayed JOLs. Lastly, we showed that delayed JOLs are not sensitive to whether they are predicting recognition or recall (so called theory-based influences) unless participants make a different prediction on each trial (i.e. trial-level design, Chapter 5). Overall, delayed JOLs are shown to vary with variables that fluctuate on a trial level, which can but do not necessarily need to map onto memory. The results suggest that delayed JOLs are primarily comparative judgments, involving the evaluation of the quantity and quality of evidence on any given trial in the context of the task at hand (e.g. by comparison to preceding trials). This is contrary to how it is often treated in the delayed JOL literature but is consistent with other metacognitive paradigms.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC	-	Area Under the Curve
FA’	-	Adjusted False Alarm Rate
FAR	-	False Alarm Rate
FOK	-	Feeling of Knowing
H’	-	Adjusted Hit Rate
HR	-	Hit Rate
JOL	-	Judgment of Learning
LSA	-	Latent Semantic Analysis
ROC	-	Receiver Operating Characteristic
SDT	-	Signal Detection Theory
SVM	-	Support Vector Machine
TLE	-	Temporal Lobe Epilepsy
TOT	-	Tip of the Tongue

CHAPTER 1

GENERAL INTRODUCTION

"The person is not a mere medium through which information flows" (Koriat, 2007)

"*The universal conscious fact is not "feelings and thoughts exist", but rather 'I think' and 'I feel'"* (James, 1890).

1.1 Introduction

The question of consciousness and self-awareness has a long history in both philosophy and psychology. The central questions revolve around the feasibility of studying self-reflective processes and the implications of such study. When Descartes asserted 'I think therefore I am' (1641/1911), he claimed in essence that his subjective awareness of his cognition was the sole certainty in which he could ground his belief in his self-hood and existence. Some have suggested that the ability to evaluate cognitive processes might even be uniquely human although that view has now been questioned (for review and further discussion see Metcalfe & Son, 2012). Either way, cognition about cognition (thinking about thinking), or metacognition, is the field that has emerged from these types of concerns and questions. Metacognition is generally defined as any judgment made about a mental event (rather than about a stimulus present in the environment). Broadly, this thesis examines how metacognitive judgments relate to the underlying cognitive processes, with a focus on memory.

At the core of metacognition is the acknowledgement that we are not merely processing machines. We do not just think or remember or imagine or perceive but we are also *aware* that we are engaging in these cognitive processes. As a consequence, we, for example, easily make the distinction between forgetting and not knowing in our daily lives (Glucksberg & McCloskey, 1981) and can distinguish a range of epistemic states (see Arango-Muñoz, 2013; de Sousa, 2009). This has important implications for our optimal cognitive function. For example, the inability to separate imagining from remembering leads to confabulations and delusions; similarly, the awareness of which of these processes we are engaging protects us from falsely concluding we are remembering when in fact experiencing novel situations (see for example Moulin, Conway, Thompson, James, & Jones, 2005). This can be seen particularly in patients with cognitive deficits who also present with anasognosia, which corresponds to a lack of awareness concerning loss of function (cognitive or otherwise; see Ernst et al., 2016; McGlynn & Schacter, 1989). Lack of awareness leads to inability to adopt appropriate compensating strategies and contributes to suboptimal (cognitive) function.

Correspondingly, cognition and awareness of cognition (metacognition) are thought to be distinct but closely related processes that depend on each other (see Nelson & Narens, 1990). That they are distinct capacities is supported by research which shows there are instances when the two can be differentially affected or even stand in opposition to each other. For example, research with clinical populations has demonstrated that it is possible to have impaired cognitive abilities and preserved metacognitive abilities (see for example Howard et al., 2010; Illman, Kemp, Souchay, Morris, & Moulin, 2016; Shimamura & Squire, 1986; Souchay, Bacon, & Danion, 2006). Every-day examples of dissociations between metacognition and cognition are déjà vu and tip-of-the-tongue (TOT) experiences. Déjà vu is the experience of familiarity combined with the awareness that the experience is misplaced and that the item or situation eliciting it are in fact novel (for a review see Brown, 2004).

TOT on the other hand is the failure to remember something from memory coupled with the knowledge that the sought-after-information is known (for a review see Brown, 2012). As such both experiences can be seen both as a failure of memory as well as a metacognitive success in that one is aware they are experiencing a memory error.

Despite these dissociations, in most instances metacognition and cognition are closely associated. Metacognition has been closely related to executive functions (e.g., attention, error correction, planning) in the literature (see Shimamura, 2000). Nevertheless, Souchay, Isingrini, and Gil (2002) observed that accuracy of metacognitive judgments about memory in Alzheimer's patients was correlated to memory performance rather than executive function scores. Further, a range of studies have shown that variables known to influence memory performance (e.g., dividing attention at study leading to shallower encoding and worse memory performance) similarly impact accuracy of metacognitive judgments about memory (Sacher, Taconnat, & Souchay, 2009).

Overall, metacognition has developed from and retains close links to memory research. To give a few examples; the distinction between memory for events (episodic memory) and memory for general knowledge information (semantic memory; Tulving, 1972, 1973) has been echoed in research on metacognitive judgments pertaining to memory (Reggev, Zuckerman, & Maril, 2011; Souchay, Moulin, Clarys, Taconnat, & Isingrini, 2007). Evidence from fMRI research has demonstrated dissociable neural substrates supporting metacognition of semantic as compared to episodic material (Reggev, Zuckerman & Maril, 2011; Elman, Klosterman, Marian, Verstaen & Shimamura, 2012). Similarly, in ageing there is evidence for preserved semantic metacognitive monitoring in instances where episodic monitoring is impaired (Morson, Moulin, & Souchay, 2015). Additionally, memory is accompanied by subjective feelings such as familiarity and some have suggested that these experiences are inferred from cues in the environment (e.g., fluency of processing; Jacoby & Whitehouse,

1989) and interpreted in the present context (Whittlesea, 1997). The same has been proposed about metacognition, which is seen primarily as an inferential process (see Koriat, 2000; Metcalfe & Dunlosky, 2008).

The central theme of this thesis is the link between memory and metamemory judgments. The focus is particularly on judgments made just after learning during consolidation, predicting future retrieval (delayed judgment-of-learning or JOL). The experimental chapters further touch on secondary questions regarding metacognitive theory and methods. The aims throughout are to (i) broaden our understanding of the processes underlying metacognitive judgments as well as (ii) to shed light on some of the methodological assumptions in the current literature and discuss their validity. This chapter introduces the background literature, starting with a general overview of research on metacognition of memory. After covering the theoretical background, the main paradigm used throughout the thesis (the delayed JOL task) is introduced and contrasted with other, similar paradigms. The rest of the chapter then focuses on issues relevant throughout the thesis, laying the groundwork and background for the different strands explored along with methodological and analytical concerns. The chapter ends with an overview of the aims of each subsequent chapter in the thesis.

1.2 Metamemory paradigms

Metacognitive research developed from and is firmly rooted in memory research and it is only more recently that it has extended to other domains. Examples of these include perception (Fleming et al., 2015; Rahnev, Koizumi, McCurdy, D'Esposito, & Lau, 2015), reasoning and decision making (Ackerman & Thompson, 2014; Fletcher & Carruthers, 2012) and agency or judgments about the sense of being in control of one's own actions (Metcalfe, Eich, & Miele, 2013; Sidarus & Haggard, 2016). Metacognitive judgments about memory are also refereed to as metamemory, a term originally introduced by Flavell (1971).

The general framework for metacognition that has prevailed to this day was introduced by Nelson and Narens (1990). It distinguishes between an object-level that represents the particular cognition under consideration and a meta-level which represents the higher-order evaluation of cognition (see Figure 1.1). As is clear from the model, the two levels are connected by two distinct processes; monitoring and control. While most early research has focused on the idea that monitoring influences control, more recent results have shown that the influence is bi-directional (Koriat, Ma'ayan, & Nussinson, 2006). This means that while control can be and often is a result of monitoring (e.g., if I feel that I have not learned something well enough I may choose to spend more time studying it), it is also possible for monitoring to result from feedback provided by control processes (e.g., if I find something difficult to process and require to increase the effort employed in studying it, I may judge that I am unlikely to remember it on some future memory test). The advantage of the Nelson and Narens model is that it provides a framework within which to study metacognitive processes that to date remains relevant. Further, it highlights both the close relationship between cognition and metacognition and their status as separate processes. This thesis focuses on monitoring exclusively and asks how participants construct judgments about memory.

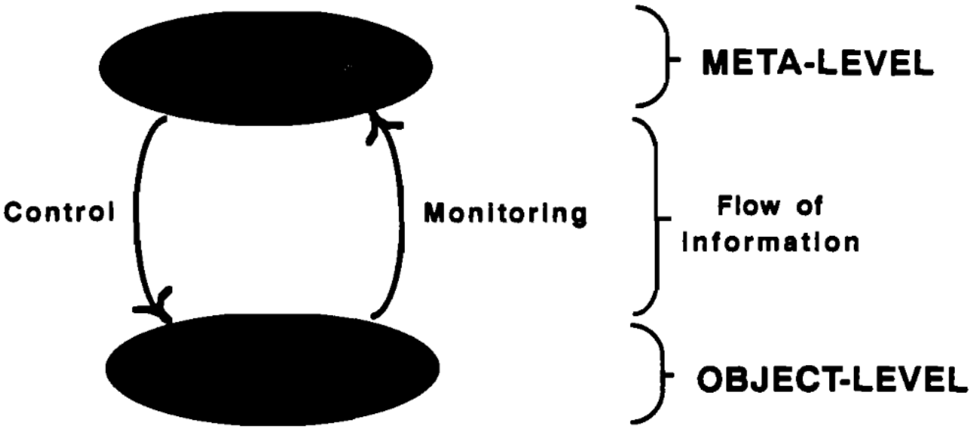

Figure 1.1: **Nelson and Narens's (1990) model of metacognition.**

Metamemory judgments can be made at any point in the memory process from prior to encoding (i.e. ease of learning judgments, assessing how easy it will be to learn the material at hand) to post-retrieval (i.e. retrospective confidence in the accuracy of what has been retrieved). From a clinical perspective, it is common to ask for a self-report of general cognitive function, either informally or through using a range of standardized cognitive questionnaires (e.g., Broadbent, Cooper, FitzGerald, & Parker, 1982; Dixon, Hultsch, & Hertzog, 1988). One can then compare the self-report to standardized tests of memory performance to assess its accuracy. However, a more sensitive approach to studying metamemory, particularly with the aim of advancing our theoretical understanding, is to ask for judgments about learning and retrieval for a particular set of items. Such an approach allows the teasing apart of specific variables that influence metacognition, variables that influence memory and variables that influence both.

As can be seen in Figure 1.2 there is a wide range of paradigms that have been employed in the metamemory literature, each tapping into different aspect of the memory process (see Nelson & Narens, 1990). While some of these judgments predict future memory performance, others retrospectively assess its accuracy. Altogether these differences mean that each judgment has a different basis and is subject to different influences. Monitoring tasks most commonly ask for judgments on an item-by-item basis (as explored in this thesis) but can also be made in aggregate, global terms (e.g., Moulin, Perfect, & Jones, 2000). The item-by-item judgment is in many ways preferable as it is easier to tease out factors that influence metacognition.

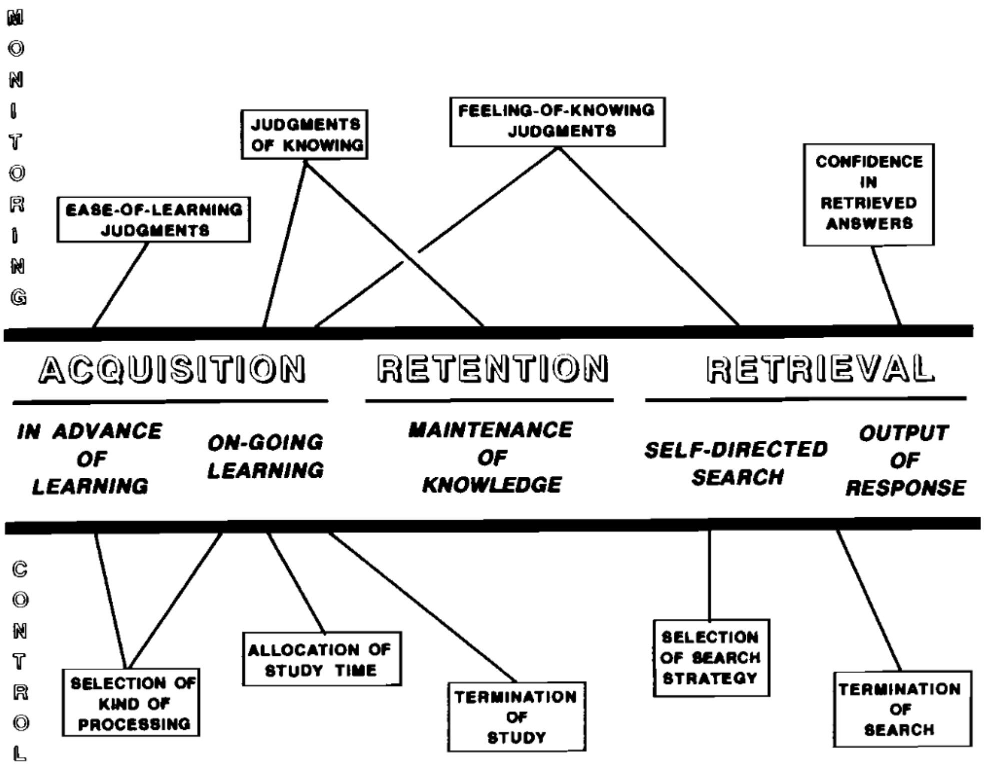

Figure 1.2: **Nelson and Narens's (1990) schematic of different types of metamemory judgments and the related memory processes.**

This thesis focuses on judgments made just after learning, during retention or consolidation. In the original classification of metamemory judgments depicted in Figure 1.2, these were called judgments of knowing but now the common term is judgment of learning (JOL; e.g., Nelson & Dunlosky, 1991). As is clear from the diagram, there is also some overlap between JOL and the feeling-of-knowing (FOK) task (Hart, 1965). The key difference is that whereas JOL tracks the acquisition and retention of information, the FOK is primarily a retrieval-oriented judgment. This means that the FOK task can be used on both material learned by participants prior to taking part in the experiment (e.g., general knowledge information) and material learned during the experiment, introduced by the experimenter (e.g., novel word-pairs such as OCEAN-TRUTH). JOLs on the other hand are studied exclusively in the context of material learned during the experiment. The most common FOK and JOL paradigms use cue-target word pairs although some studies have also used images as cues and/or targets (e.g., Chua, Hannula, & Ranganath, 2012; Metcalfe & Finn, 2008b). All experiments reported in the present thesis employed cue-target word-pairs.

1.2.1 The delayed JOL

The JOL task has a number of strikingly different formats in the literature. The greatest difference is between the immediate and delayed JOL paradigm (Nelson & Dunlosky, 1991). In immediate JOLs, participants study the cue-target pairs one at a time and after the study of each pair, they make a judgment about whether they will retrieve the target when presented with the cue on a memory test which is administered after all items have been studied and judged (e.g., Hanczakowski, Zawadzka, Pasek, & Higham, 2013; Koriat, 1997; Rhodes & Castel, 2008). In this form it is very much a judgment tracking acquisition. In the delayed JOL paradigm, there is a delay between when each pair is studied and when a JOL for that pair is given. In the most common form of the delayed JOL, participants first study *all* cue-target pairs without making any judgments. Only after all items have been studied are participants presented with all the studied cues, one at a time, and are asked to make a prediction about their confidence in their ability to retrieve the target on the subsequent memory test (e.g., Metcalfe & Finn, 2008b). Delayed JOLs are usually significantly more accurate at predicting performance than immediate JOLs although this difference disappears if the delayed JOL format employs both the cue and the target at judgment (Dunlosky & Nelson, 1997). Much research has focused on trying to understand the difference in accuracy between the two judgments. While a consensus hasn’t been reached yet, all the existing theories focus on differences in the basis of the two judgments, arguing the cues informing delayed JOLs are more indicative of future memory performance than those informing immediate JOLs (for a review see Rhodes & Tauber, 2011). As such the two judgments are considered very different in terms of the influences they are sensitive to and their underlying processes (see also Koriat, 1997).

In contrast, there are clear similarities between the delayed JOL and the FOK task. Classically, the FOK requires participants to first attempt recall of the target item and only in instances when they cannot retrieve it, are participants asked to judge whether they feel they know the target enough to recognise it on a subsequent recognition test (e.g., Hart, 1965; Jersakova, Souchay, & Allen, 2015; Souchay & Isingrini, 2012). While traditionally only trials on which the target was not retrieved were considered, more recently it became common to ask for an FOK judgment for all items irrespective of whether the target was retrieved. The latter, while now usually called FOK, has also been termed a prediction of knowing (POK; Schreiber & Nelson, 1998).

In many ways then the delayed JOL, FOK and POK are very closely related judgments. In particular, the recent extension of FOK to all trials rather than just failed recall trials increases its similarity to a delayed JOL. As a result the difference between a delayed JOL and an FOK is primarily in that the latter requires an overt target retrieval attempt whereas in the delayed JOL it is implicitly assumed participants attempt retrieval in making the judgment but it is not explicitly required of them. Nevertheless, a study has shown that response latencies for JOLs given without first attempting to retrieve the target (overtly or covertly) are different to response latencies for JOLs where participants were not explicitly instructed to retrieve the target (Son & Metcalfe, 2000). More specifically, in the first case JOL magnitude was inversely correlated with response latencies whereas in the latter case the relationship between JOL magnitude and response times was a U-shaped function. Further, it was observed that judgment accuracy on episodic delayed JOL and FOK tasks was not correlated and that whereas FOK accuracy correlated with executive function measures, JOL accuracy did not (Souchay, Insingrini, Clarys, Taconnat, & Eustache, 2004). Further, older adults were shown to be impaired on FOK accuracy but not on delayed JOL accuracy (Souchay & Isingrini, 2012). This means that findings from the FOK literature do not necessarily generalise to delayed JOLs and despite their clear similarities, the two tasks need to be considered separately.

Whether researchers have employed FOK or delayed JOL has often been guided by the key questions of interest. The FOK literature is primarily concerned with the type of information and manipulations that influence the judgment (e.g., Koriat & Levy-Sadot, 2001; Koriat, 1993; Metcalfe, Schwartz, & Joaquim, 1993). This has likewise been the primary focus of immediate JOLs (e.g., Rhodes & Castel, 2008). The delayed JOL literature on the other hand has primarily focused on JOL accuracy with a particular focus on attempting to answer why delayed JOLs are more accurate than immediate JOLs (for reviews see for example Metcalfe & Dunlosky, 2008; Rhodes & Tauber, 2011). By focusing on the delayed JOL as a special case of the immediate JOL, the understanding of the underlying processes guiding the delayed JOL are somewhat limited. There have been only a few studies that have attempted to understand the basis of the delayed JOL in more detail and in its own right (e.g., Metcalfe & Finn, 2008b) and the delayed JOL literature is in this aspect less rich than the work pertaining to the other paradigms. The primary goal of this thesis is to add to this sparser literature by investigating the underlying mechanisms that drive delayed JOLs.

1.2.2 Overview of experimental paradigm

The focus of this thesis is on the underlying processes of delayed JOLs. More specifically, our focus is on how participants generally assess access to recently learned information and predict the likelihood of future retrieval for that information. In this thesis we employed a classic delayed JOL task throughout with only the cue presented at the Judgment Phase, which was administered only once all items have been studied. The criterion memory task for

JOLs is usually cued recall (e.g., Benjamin, 2005; Metcalfe & Finn, 2008a, 2008b) but can also be recognition (e.g., Dunlosky & Nelson, 1997). Majority of the chapters here used recognition as the criterion task but we also used in recall in Chapter 5.

To give an overview of the basic paradigm used throughout the thesis (see Figure 1.3), it consists of three distinct phases (with modifications or additions of further phases in some chapters). Firstly, the Study Phase presents participants with a list of cue-target word pairs, one at a time. This is followed by a Judgement Phase where participants are presented with the cue of each studied pair one at a time. They are asked to indicate whether they will retrieve the target on the subsequent memory test. These judgments are either made on a confidence scale (0%, 20%, 40%, 60%, 80%, 100%) or as a binary (*yes/no*) judgment. Lastly, in the Memory Phase participants are presented with all studied cues along with a number of options from which to choose the target corresponding to the presented cue. All options are targets that have been studied. Overall, delayed JOLs have been shown to accurately map onto memory performance with higher JOLs given for remembered as compared to not remember targets (for review on delayed JOL accuracy see Rhodes & Tauber, 2011).

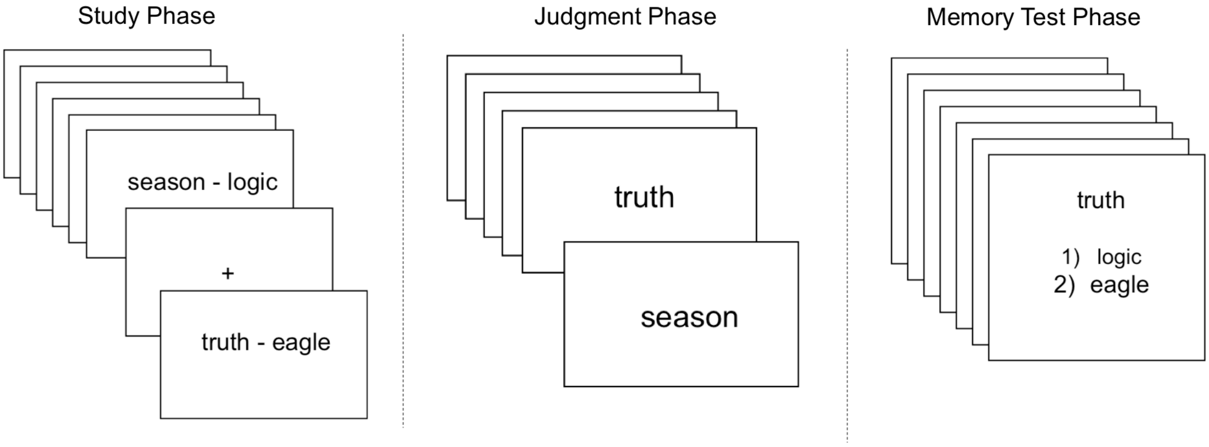

Figure 1.3: Schematic of general procedure.

1.3 Metamemory theories

One of the biggest questions in metacognitive research is how metacognitive judgments are constructed. What are the types of influences and information that participants incorporate into the judgments they make and which of these influences lead to optimal judgments and which distort them? Broadly, a common factor underlying all metamemory theories is that they describe metamemory judgments in relation to the underlying memory processes, although the view of the precise link has changed over time.

1.3.1 Early metamemory theories

The first account of metamemory, primarily stemming from FOK and TOT research, was that of direct partial access to the target item (Eysenck, 1979). This presupposed that even in instances when the target could not be fully retrieved, one had access to the target features and the target trace was at least partially activated; the stronger the partial access the higher the feeling that one knows the sought after information and that it would subsequently be retrieved in the near future. This is based on a wealth of research which has demonstrated that when people cannot fully recall the target item, they can still recall partial information about it such as the first letter and other orthographic information or semantic content (for review see Brown, 2012). When participants can access this partial information they also give higher FOK ratings and are more likely to report being in a TOT state, indicating they are on the verge of target recall. Similarly, when they report a TOT state or a high FOK, participants are also more likely to retrieve the target item (Gardiner, Craik, & Bleasdale, 1973). The notable aspect of this theory is that of a *privileged* access to the memory trace of the sought-after information. However, it was also noticed when participants remembered *incorrect* partial information, they likewise reported higher FOKs (Koriat, 1993; Thomas, Bulevich, & Dubois, 2012). This undermined the idea of privileged access to the target trace and rather suggested that the judgments might be inferred from *any* partial information coming to mind (irrespective of accuracy).

From these observations the target accessibility account developed. This states that participants monitor the level of *perceived* access to the target item through *cues* such as retrieval of partial information or ease of target retrieval at time of judgment. It is then the quantity as well as the intensity of the retrieved information that informs the metacognitive judgment. This account is different to the one mentioned above in that it is not assumed that participants have privileged, direct access to the target item. Because the information that comes to mind is usually mostly relevant and target related, judgments are overall fairly accurate (Koriat, 2000). Nevertheless, any information that comes to mind and that appears relevant at the time of judgment can inform that judgment.

Another early account focused on the cue rather than the target. More specifically, it has been observed that metacognitive judgments increase with experimentally manipulated familiarity with the cue term (Metcalfe, Schwartz, & Joaquim, 1993; Reder & Ritter, 1992; Reder, 1987). The most common method of manipulating cue familiarity is to expose participants to some items in a seemingly unrelated task prior to the Study Phase (e.g., Liu, Su, Xu, & Chan, 2007). An example would be a pleasantness-rating task where participants are presented with one item at a time and asked to rate its pleasantness on a scale. Following this they would complete a standard Study-Judgment-Memory Test paradigm with the cue of half of the studied cue-target pairs having been seen in the rating task. Participants give higher i.e. more confident metacognitive judgments when presented with a familiar cue across a range of paradigms. Some studies have found that manipulating cue familiarity increases access to the target, indexed by increased likelihood of retrieval (Liu et al., 2007; Metcalfe & Finn, 2008b) while other studies have not observed this (Benjamin, 2005). It makes sense that familiar information should also be better known, but the varied memory results mean the debate continues as to whether cue familiarity is a diagnostic cue.

1.3.2 Modern metamemory theories

Current models of metamemory highlight the *combined* contributions of cue familiarity and target accessibility to metacognitive judgments as opposed to focusing on only one source of influence on metacognitive judgments. The first to propose this view were Koriat & Levy-Sadot (2001) who outlined a two-stage model of FOK judgments. This was later also extended to delayed JOLs (Benjamin, 2005; Metcalfe & Finn, 2008b). The combined view not only reconciles cue familiarity and target accessibility as both contributing to metacognitive judgments but also suggests how the two effects interact. In this view, metamemory judgments are described as a two-stage process. The first stage is fast acting and driven by cue familiarity; if the cue does not feel familiar, then a negative or a low confidence judgment is made outright without engaging stage two. If the cue feels familiar then stage two, which comprises of a target accessibility assessment, is initiated. It makes sense that if the cue does not feel familiar there is no point in searching for its associated target, a process which is slow and effortful. The cue thus provides a quick signalling system for deciding when and how to engage target retrieval mechanisms.

An important shift in the literature from the original, direct access view is that these models are inferential in nature. In other words, the general idea is that when making a metacognitive judgment, such as whether a target is or will be accessible, a range of subjective cues and feelings are considered (Koriat, 2000). The strength of the experiences that arise (e.g., the experience of familiarity with the cue, or ease of access to the target) are used to *infer* the metacognitive judgment given. This is also why cue familiarity and target accessibility are sometimes also referred to as experience-based influences. The inferential view of metacognition remains the most prominent and generalizes across metacognitive paradigms. For example, fluency of stimulus processing has been shown to influence judgments made about memory (Rhodes & Castel, 2008) as well as judgments made about agency (Sidarus & Haggard, 2016) and reasoning (Thompson et al., 2013).

Within this framework, the metamemory literature has also recently begun to examine what other sources of information, beyond that of cue familiarity and target accessibility, influence metacognitive judgments. The noncriterial recollection hypothesis stems from results showing that memory for one source dimension influenced judgments made about access to another source dimension (Brewer, Marsh, Clark-Foos, & Meeks, 2010). Further, access to encoding strategies, namely what link was made between the cue and the target at study, influences judgments about access to the target (Hertzog, Fulton, Sinclair, & Dunlosky, 2014). Altogether, this new line of evidence (to date only investigated in relation to FOK judgments) emphasizes metacognition as a general evidence evaluation system where evidence refers to any seemingly related information that comes to mind at the time of judgment. Some of these ideas are considered in more detail in Chapter 2.

1.3.3 Types and levels of metacognitive influences

Additionally, throughout the literature, researchers have often drawn a distinction between different *types* of influences on metacognitive judgments. Starting with Flavell and Wellman (1977), they distinguished between person, task and strategy-related influences, which they termed variables. All of these refer to knowledge that participants have about learning and memory. Person variables refer to, for example, knowledge of how well one learns, task variables refer to knowledge of how task manipulations (e.g., delay between learning and recall) affect retrieval, and strategy variables refer to knowledge about the usefulness and effectiveness of different learning strategies. Koriat (1997), focusing on judgments of learning, made a distinction between intrinsic, extrinsic and mnemonic influences (or as he referred to them – cues). Intrinsic cues are for example characteristics of the studied items (e.g., level of relatedness between the cue and the target), extrinsic cues refer more specifically to the entire learning situation (e.g., time given to study each item), and mnemonic cues are considered internal and subjective (e.g., feeling of knowing the sought after target information). More recently Koriat (2000; see also Arango-Muñoz, 2010) suggested that the distinction could be made between implicit and explicit influences. Explicit (or judgment-based) influences collectively refer to what Koriat earlier called intrinsic and extrinsic influences. In other words, judgment based influences can be described as general knowledge and beliefs one has about learning and memory. These beliefs also encompass all variables originally described by Flavell and Wellman (1977). Mnemonic cues are thought of as implicit and heuristics-based. Heuristic in this case refers to the idea that they are not necessarily accurate, consistent with the inferential account of metamemory described above. When I am trying to recall the name of a book I read last month I might feel that the title started with the letter *S* while in fact the title was *Americanah.* However, based on the fact that any partial information comes to mind as I am thinking of the name, I *infer* that I am about to remember it. Cue familiarity and target accessibility are considered both to be examples of implicit heuristics that drive metacognitive judgments. The majority of this thesis (Chapters 2, 3, and 4) is focused on investigating implicit heuristics.

While the distinction between implicit and explicit influences on metacognitive judgments has remained popular, explicit influences have been less researched and evidence for their impact on judgments remains limited. The influence of explicit processes on metamemory judgments has been most clearly examined using immediate JOLs. Variables such as the semantic relationship between the cue and the target (Castel, McCabe, & Roediger, 2007), how far in advance the predicted memory test is to take place (e.g., in 10 minutes as compared to in a week; Koriat, Bjork, Sheffer, & Bar, 2004) and how many learning opportunities will be given for each cue-target pair (Kornell & Bjork, 2009) all influence JOL predictions. However, it is noteworthy that most commonly these variables are manipulated on a trial-level. As such, for example, a participant might encounter a semantically related cue-target pair on the first trial and a semantically unrelated cue-target pair on the second trial. Similarly, they might predict target retrieval in 10 minutes on a given trial and retrieval in a week on the subsequent trial. Those studies that have contrasted these trial-level designs with blocked or between-subject designs (Koriat et al., 2004; Kornell & Bjork, 2009) have found that the effect of these explicit types of variables on metamemory judgments disappears. Altogether, these studies provide converging evidence that participants do not actively use theory-based, explicit processes in making online judgments such as immediate JOLs even though when asked directly, participants do hold such theories about memory (Kornell & Bjork, 2009; van Velzen, 2013). Delayed JOLs have been found to change with whether participants are predicting future recognition or recall such that participants are more confident when predicting recognition as compared to recall, though so far this has only been examined in a trial-level design (Mazzoni & Cornoldi, 1993; Thiede, 1996). As such the role of explicit influences in metamemory judgments remains less clear. This is discussed in more detail in Chapter 5.

1.4 Measures of metacognitive accuracy

Before outlining the aims of the thesis as a whole, it is necessary to make some comments on the general analysis approach to the data in all the experimental chapters. Metacognitive accuracy is traditionally analysed either in absolute (calibration) terms or in relative (resolution) terms. Absolute accuracy looks at overall correspondence between metacognitive judgments and cognitive performance (e.g. whether for all items given a 60% JOL participants remembered 60% of those items). If such correspondence between average performance and percentage JOLs given were true for all JOL responses, then a participant would be considered perfectly calibrated. Correspondingly, if participants overall gave 50% of *yes* JOL predictions, they would be perfectly calibrated if they remembered 50% of all items. If participants' performance is higher than their JOLs, they are said to be underconfident whereas if it is lower they are overconfident. Much research has focused on poor calibration across tasks and populations (e.g., Kelley & Sahakyan, 2003). The underlying assumption of calibration however is that participants use confidence in probabilistic terms and recent research has suggested this might not be the case (discussed in more detail in subsequent chapters, see Hanczakowski et al., 2013). Another issue is that the analysis ignores item-by-item correspondence between judgments and performance, which is the focus of relative accuracy measures. The primary focus of this thesis thus is on relative accuracy.

The most straightforward way to assess relative correspondence is to use correlations and the most commonly employed (see Nelson, 1984) in metacognitive tasks is the Goodman-Kruskal gamma correlation (Goodman & Kruskal, 1954). Gamma compares the judgment and memory outcome for each item against all other items. If an item that is subsequently recognised receives a higher judgment than an item that is not recognised, that constitutes a concordant pair. If the recognised item receives a lower judgment than the not recognised item then that is a discordant pair. Gamma is based on comparing the number of concordant and discordant pairs (ignoring ties). The fact that each item is compared to all other items is what makes it a relative measure of accuracy.

An alternative approach to relative accuracy employed throughout this thesis is the signal detection theory (SDT) approach. The advantage of this approach lies in that it can separate influences of participants' response bias (criterion; their tendency to endorse one response more than another) and their ability to discriminate between the classes of interest (see Macmillan & Creelman, 2004). When the classes refer to stimulus specific information (e.g., is it *old* or *new*; is it *present* or *absent*) we refer to type-1 SDT. When the classification refers to accuracy of responses, and metacognitive judgments in general, it is referred to as type-2 SDT (see Higham, 2011). Naturally, the focus here is on type-2 SDT analysis as a method for evaluating accuracy of metacognitive judgments.

Relative accuracy or discrimination is commonly assessed using *d*-prime (abbreviated *d*') for binary (*yes/no*) judgments. This measure assumes that participants are evaluating two signals (one providing evidence for future target retrieval and one providing evidence to the contrary) in the presence of noise. The two signals are thought to be normally distributed and with equal variances. The distance between the two distributions corresponds to the participant's ability to distinguish targets that will be retrieved from those that will not and is captured by *d'*.

To calculate *d'* for each participant we calculate the number of hits (*Hn*; number of remembered items given a *yes* JOL prediction), misses (*Mn*; number of remembered items given a *no* JOL prediction), correct rejections (*CRn*; number of not remembered items given a *no* JOL prediction) and false alarms (*FAn*; number of not remembered items given a *yes* JOL prediction). Using a Snodgrass & Corwin (1988) correction we then calculate the adjusted hit rate (*H'*) and adjusted false alarm rate (*FA'*) which are then used to calculate *d'*:

$$(1)\ H' = \frac{Hn + 0.5}{Hn + Mn + 1}$$

$$(2)\ FA' = \frac{FAn + 0.5}{FAn + CRn + 1}$$

$$(3)\ d' = z(H') - z(FA')$$

The Snodgrass and Corwin correction consists of adding 0.5 to the numerator and 1 to the denominator and was developed to deal with ceiling and floor effects. For example, if a participant had a perfect hit rate then without the correction it would not be possible to obtain *z*-values necessary for calculating *d'*. The *d'* values usually range from 0 to 3, where 0 indicates a full overlap between the two distributions and a corresponding inability to discriminate between the two classes of interest (here items that will and will not be remembered). The higher the value of *d'*, the higher the accuracy of the metacognitive predictions under investigation.

Some researchers have also calculated *d'* from confidence data by splitting the JOL confidence scale into assumed *yes* and *no* predictions (e.g., <=40% JOL confidence means a *no* JOL prediction and >=60% JOL confidence means a *yes* prediction; see for example Mason & Rottello, 2009). However, this makes assumptions about how participants interpret the confidence scale, which seems unlikely to be interpreted by all participants in all contexts in this manner (see Chapter 4). Rather, each participant can set their own threshold for negative and positive responding and this is likely to vary across experiments (see for example Serra & England, 2012).

Due to these considerations, it is more appropriate to use the Area Under the Curve (AUC) measure of resolution for confidence responses (a commonly employed SDT measure of confidence accuracy). To compute AUC, 20%, 40%, 60%, 80% and 100% JOLs are each in turn considered a response threshold for a *yes* JOL prediction (starting with 20% and above = *yes* JOL). For each threshold, the corresponding hit rate (*HR*) and false alarm rate (*FAR*) are calculated. These are in essence the same as *H'* and *FA'* calculated using equations 1 and 2 but without the Snodgrass and Corwin correction which in this case is not necessary. The results are then plotted against each other with *FAR* on the *x*-axis and *HR* on the *y*-axis starting with the highest, most conservative response threshold (here, 100% = *yes*) in the bottom left corner of the graph. This results in a curve called the Receiver Operating Characteristics (ROC) curve (see Figure 1.4 for an example of such a curve). AUC is the area under this curve calculated with a trapezoidal rule. In contrast to *d',* AUC ranges only from 0-1 but is interpreted in the same way; the higher the value, the better the resolution.

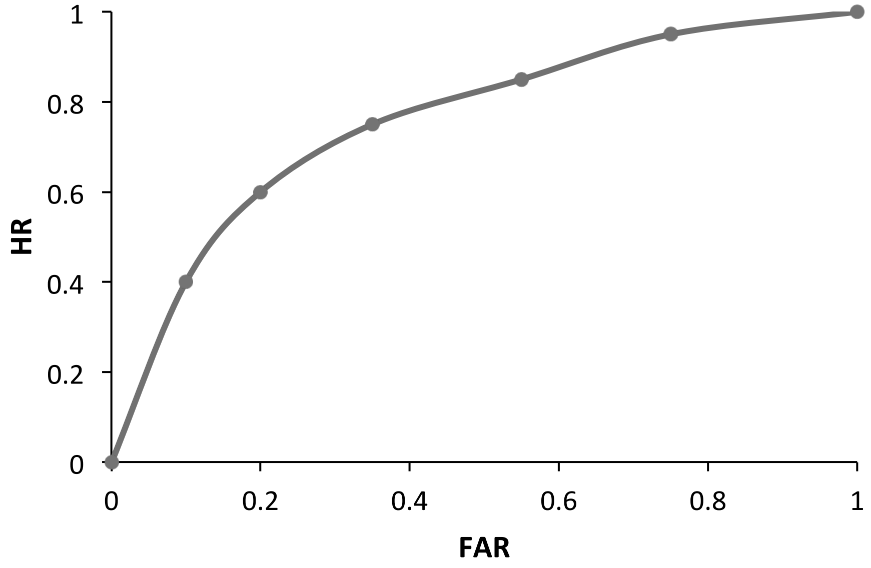

Figure 1.4: Example ROC curve. *Each point represents different confidence criteria. The most conservative criterion (100%) is plotted first (left bottom corner) and each subsequent point represents a more liberal criterion. HR = Hit rate, FAR = False Alarm rate.*

Lastly, under the SDT model, it is possible for two participants to be equally accurate even though one is more likely to give a *yes* JOL prediction than the other (for example). The response bias-free nature of *d'* and AUC as measures of accuracy is one of the great advantages of the SDT approach. Further, analysing response bias (or criterion) is a useful way to further unravel participant responding across conditions. It is possible to calculate criterion (abbreviated c), from binary responses using hit rate (*H'*, equation 1) and false alarm rate (*FA'*, equation 2), using the following equation:

$$(4)\ c = -\frac{1}{2} * [z(H') + z(FA')]$$

A criterion of 0 denotes no response bias. Positive criterion (values above 0) is a conservative response bias or a greater tendency toward negative (i.e. *no*) predictions whereas a negative criterion (values below 0) indicates a liberal response bias (greater tendency toward *yes* predictions).

The clear advantage of the SDT approach is that one can separate accuracy from response bias. The ability to discriminate between two signals (e.g., what will be remembered vs. what will not be remembered) and individual tendency toward responding in a specific way are clearly two distinct phenomena. It is for this reason that throughout this thesis we employ *d'* and AUC to assess relative accuracy. In Chapters 2 and 5 where we also investigated participant responding we further looked at bias in participant responding. Altogether, the SDT approach thus gives us increased sensitivity in investigating the effects under investigation. The use of SDT analyses while employed in the literature (e.g., Hanczakowski et al., 2013; Higham, 2011; Zawadzka & Higham, 2016), is still not very common and as such this is one of the novel contributions of this thesis.

1.5 Thesis aims

In summary, this thesis focuses on metacognitive monitoring just after acquisition and during retention, examining the processes underlying judgments predicting future item retrieval. The method employed is the delayed JOL paradigm using cue-target word-pairs. While the FOK and immediate JOL literature has examined in detail what drives these judgments, the delayed JOL literature has done less in this respect. Despite the overlaps between the various metamemory paradigms, there are clear differences between them and the underlying memory processes they relate to. It is for this reason that a more detailed analysis of the determinants of delayed JOLs is warranted. The particular focus of this thesis is on how the delayed JOL relates to the underlying memory processes.

Episodic memory is seen as a collection of separate features and attributes. This is true in the case of memory for simple, single items as much as in the case of memory for complex, composite scenes (Horner & Burgess, 2013; Tulving & Thomson, 1973). The partial access and inferential views of metamemory are clearly based around this broad idea that aspects of the memory event can be accessed even when the full representations cannot be brought to mind. Related to this is the view that information stored in memory can be available even when it is not fully accessible, as indicated by the ability to recognise items that cannot be recalled (Tulving & Pearlstone, 1966). This is the foundation of metamemory theories to date.

The significance of the move towards inferential theories of metamemory is that they no longer suppose a direct, privileged link between memory and metamemory. In other words, the information that one brings to mind about the cue-target pair being judged in any metacognitive task is likely to be correct but does not need to be; one could bring to mind information completely unrelated to it. This is why the judgment is described as being inferred from the cues and information available at time of judgment.

Correspondingly, it is often the case that variables that influence memory also influence metamemory, but the influences can be dissociated (Schacter, 1983). Manipulations such as dividing attention at study have been found to decrease both memory performance and FOK accuracy (Sacher et al., 2009). On the other hand, while target priming has been shown to improve target memory it did not influence FOK judgments (Jameson, Narens, Goldfarb, & Nelson, 1990). Similarly, immediate JOLs have been shown to be subject to a range of metacognitive illusions such as judgments changing with font size of the cue-target pairs, with font size having no corresponding effect on memory performance (Rhodes & Castel, 2008).

In contrast, delayed JOL theories have mostly assumed that the judgment is based on evaluating access to the target. This is a result of the delayed JOL literature having focused primarily on explaining why it is more accurate than the immediate JOL; the uniting thread of all current theories being that the delayed JOL is based on recallability of the target at time of judgment (Rhodes & Tauber, 2011). The only extension to that view has been the suggestion that cue familiarity can also play an initial although limited role in guiding delayed JOLs (Metcalfe & Finn, 2008b). As such the aim of the present thesis is to ameliorate our understanding of the processes underlying delayed JOLs by focusing on variables related to memory and the idea of metacognitive judgments as a general inferential mechanism relying on varied sources of evidence. This is explored in studies with healthy adults, using a range of experimental manipulations and methods.

More specifically, the first two experimental chapters focus on variables that might affect both memory performance and delayed JOLs. Chapter 2 extends the noncriterial recollection hypothesis (Brewer et al., 2010) to delayed JOLs. More particularly, it focuses on how access to where the target appeared at study (i.e. target source information) influences confidence that the target will be accessed (i.e. delayed JOL) and explores the boundary conditions on this effect. This is done in the context of looking at how access to this spatial information relates to memory for the target itself. Chapter 3 then turns to the classic cue familiarity effect, which has been extended to JOLs only more recently (Metcalfe & Finn, 2008b). More specifically, the chapter investigates how manipulating familiarity experimentally (as is commonly done in metamemory paradigms) compares to effects of pre-experimental familiarity. This is achieved through using cues participants encounter in their daily lives as well as cues that they have never encountered before. Again, both types of familiarity are investigated in terms of how they affect memory performance and how they influence JOL responding.

In contrast, Chapter 4 asks participants to justify their JOLs (without any further guiding instructions) and in this way explores what type of information they rely on in constructing their justifications. As with the previous chapters, the main focus is on what memory-related information (e.g., cue familiarity) participants describe as informing their judgments. However, this time this is derived from an in-depth analysis of participants' report rather than experimental manipulations of the task. The analyses also compare the features of the content of the verbal reports that dissociate different types of JOLs (e.g., 0% vs. 20% confidence or *yes* vs. *no* JOL) and compare the results to some theoretical models of how differences sources of information underlie different types of judgments (e.g., lack of cue familiarity = *no* JOL).

Whereas Chapters 2 to 4 examine the effect of implicit, mnemonic cues, Chapter 5 turns to the question of explicit, theory-driven influences. More specifically, it examines whether delayed JOLs are sensitive to the retrieval test (recognition vs. recall) they are predicting. In other words, the chapter explores whether participants would incorporate their theories about memory (recognition is easier than recall) into their prediction of the likelihood of retrieving the target.

Further, whereas Chapter 2 uses confidence judgments, Chapter 3 uses binary (*yes/no*) JOL predictions. As noted earlier, not only is there a range of metacognitive paradigms, each can be implemented in a variety of ways. To this date it has been common to use various formats interchangeably and comparisons between them have been rare. Despite this, a recent study found that confidence and binary immediate JOLs might not be equivalent, at least in immediate JOLs (Hanczakowski et al., 2013). More specifically, whereas confidence judgments were found to be underconfident in terms of calibration, binary judgments were well calibrated. This suggests either that participants employ the two response formats differently or we might be misinterpreting what confidence represents. Chapters 4 and 5 compared the two response formats directly. As such not only does Chapter 4 introduce a completely novel methodology, both Chapter 4 and 5 further touch on methodological issues only now coming to light in the metacognitive literature. Previous research has found that question framing influences responding on cognitive (Mill & O'Connor, 2014) and metacognitive tasks (Finn, 2008). However, the idea that the response options given to participants could have the same effect is fairly novel (Jersakova, Moulin, & O'Connor, 2016; Hanczakowski et al., 2013) and further explored in this thesis.

In summary, the focus of this thesis is primarily theoretical. I explore how delayed JOLs relate to memory through (i) manipulating variables and observing how they impact memory and metamemory, (ii) asking participants to tell us how they construct their judgments and (iii) investigating whether participants' delayed JOLs are sensitive to theories about memory. All metacognitive accuracy data is analysed using SDT. Further, methodological concerns are addressed by comparing response formats directly which allowed us to (i) establish the generalizability of findings, (ii) explore assumptions implicit in current interpretations of data in the literature and (iii) shed further light on how people evaluate their cognition.

CHAPTER 2

MEMORY AND METAMEMORY FOR VERBAL-SPATIAL ASSOCIATIONS

2.1 Introduction

The current theories of metamemory monitoring stress that judgments are based on the quantity of primarily target-related information that is accessible at the time of judgment. While the focus has usually been on semantic and ortographic target information (e.g., Koriat, Levy-Sadot, Edry, & de Marcas, 2003; Koriat, 1993), more recently it has been shown that other types of information access can influence FOK judgments (Brewer et al., 2010; Hertzog et al., 2014). It has also been shown that not just quantity but also quality (i.e. accuracy) of the accessed partial information can influence FOKs (Norman et al., 2016; Thomas et al., 2012). To date, neither of these ideas has been considered in the context of delayed JOLs. As such this chapter addresses (i) whether access to seemingly irrelevant target related information (its position at study) and (ii) the accuracy of that access (i.e. is the correct position remembered) influence delayed JOLs. Further, we asked how access to spatial information about the target relates to memory for the target item.

Our daily experience is grounded in a spatiotemporal context and, consequently, when and where we have encountered information forms a large part of our learning. Correspondingly, episodic memory has been defined as retrieval grounded in space and time (Henson & Gagnepain, 2010; Tulving, 1985) and some researchers have emphasized spatial context as fundamental to episodic memory (Burgess, Becker, King, & Keefe, 2001; Maguire & Mullally, 2013; Robin, Wynn, & Moscovitch, 2016). Supporting this idea is ample neuroscientific evidence for connections between episodic item memory and spatial as well as relational processing (Burgess et al., 2001; Konkel & Cohen, 2009; Moscovitch et al., 2005). Similarly, behavioural data support the idea that there are more similarities than differences between the processes that support encoding of spatial and (object) identity information (Köhler, Moscovitch, & Melo, 2001). Surprisingly, not much is known about how access to spatial context relates to item retrieval and metamemory retrieval predictions. The present chapter addresses this considerable gap in the literature, exploring whether and how memory and metamemory for verbal associations is influenced by access to information concerning where in space it was encountered. More specifically, this chapter investigates how access to information about the spatial configuration of previously learned visually presented cue-target word pairs relates to (i) memory retrieval accuracy for the learned target and (ii) metamemory judgments predicting whether the target would be retrieved.

The first aim of the study was to explore whether memory for spatial information is related to recognition memory for the target. Identity-spatial associations have been considered in the context of working-memory (see Allen, 2015) and episodic long-term memory, with research in the latter case often focusing on the role of the hippocampus in spatial and relational processing (e.g., Burgess, Maguire, & Keefe, 2002; Konkel & Cohen, 2009). It has been shown that spatial information is bound to the item while not necessarily integrated with other source information (Starns & Hicks, 2008) and that this binding occurs at initial encoding (Allen, Vargha-Khadem, & Baddeley, 2014; Uncapher, Otten, & Rugg, 2006). In contrast, the present study primarily focused on exploring processes operating at the retrieval stage, and investigating how probability of retrieving item information (what was studied) relates to probability of retrieving spatial information (where it was presented at study). This has bearing on current attempts at mapping retrieval dependency between learned items and their constituent features (Horner & Burgess, 2013, 2014; Starns & Hicks, 2005; Trinkler, King, Spiers, & Burgess, 2006). Further, and more crucially for the current study, understanding this has implications for interpreting the impact of (spatial) information access on metamemory judgments.

Research on how metamemory judgments are constructed has shown that it is largely an evidence accumulation process relying on a number of access heuristics. Overall, the consensus is that the more partial or related information that is retrieved at time of judgment (e.g., semantic or orthographic information), irrespective of its accuracy (Koriat, 2000; Koriat, 1993), and irrespective of its relationship to recognition memory for the target (Alban & Kelley, 2013; Rhodes & Castel, 2008) the more confident participants are that they will know the target. More recently, research has started asking how accuracy of access (not just its quantity) contributes to metamemory monitoring (e.g., Dunlosky, Rawson, & Middleton, 2005). Thomas et al. (2012) observed that accuracy of accessed partial information can in some instances also be a contributor to metamemory judgments. More specifically, participants were more likely to predict they will know the target (indicated by higher judgment magnitude) when the retrieved target-related information was accurate but only when the accessed information was conceptual (target category) and not perceptual (font colour) in nature. These findings might therefore suggest a distinction between perceived access (i.e. retrieving any information while searching for the target), which always influences metamemory judgments, and the accuracy of that access whose contribution to metamemory judgments is more selective. Thomas et al. (2012) specified that accuracy of retrieving partial information might only be a contributor to metamemory confidence when that information is tied to the meaning of the to-be-retrieved item (i.e. when it is conceptual). However, this conclusion is based on analysis of only very superficial perceptual features and the authors fail to specify why conceptual information should hold a special status other than that it is ‘inherent to verbal material’. If there are features that are not conceptual but fall into the visuospatial domain whose accuracy of access contributes to metamemory predictions, then spatial information seems like a good candidate given its apparent importance to episodic memory. This study offers a follow-up on the Thomas et al. (2012) findings by examining whether accuracy of non-conceptual feature access can contribute to metamemory judgments.

Across three experiments, we manipulated where the target appeared at study. After this Study Phase, participants were presented with the cue of each pair, and asked to give a JOL indicating on a scale (0-100%) how confident they are that they would recognize the target on the following memory test. Within this JOL stage we also asked participants to indicate where they thought the target appeared at study by selecting one of a number of possible locations on the screen and to indicate their retrospective confidence in the accuracy of their memory for the target location (0-100%). This was followed by a recognition test for the cue-target pairings.

In summary, the present study was designed to explore the relationship between identity and spatial location from both memory and metamemory perspectives. Firstly, we were interested in observing whether remembering where the target appeared would be related to memory for what the target was as indicated by target recognition performance. Secondly, we wanted to understand how JOLs are constructed through investigating how they relate to different types of target-related information access. Specifically, we explored whether, in addition to target recognition, JOLs would also be related to participants' retrospective confidence in whether they remembered the target location correctly (perceived access to location information) and the accuracy of their spatial memory. To capture the level to which these three variables relate to JOL magnitude and to account for individual differences in JOL responding (e.g., some participants giving overall higher JOLs), we carried out an item-level regression analysis (described in more detail in the results section). This allowed us to determine whether there is an independent significant contribution of all examined variables to JOL magnitude. In line with the broad accessibility view, we predicted that JOLs would increase with participants' confidence that they correctly identified where the target was presented at study (i.e. perceived access). We further hypothesized that this effect would be independent of whether accuracy of spatial access also contributed to JOL magnitude and whether spatial access was related to item access.

2.2 Experiment 1

The key aim of Experiment 1 was to establish whether and how access to spatial information relates to item memory and impacts JOLs. As a secondary manipulation, we also manipulated whether participants were told they would be tested on their memory for target location, to check for whether outcomes are independent of intentionality of encoding and task expectations. Past research has shown that whether participants are told to specifically encode certain information can impact memory retrieval for that information (Eagle & Leiter, 1964; Naveh-benjamin, 1987; Williams, 2010). This allowed us to determine whether participants' intention to encode spatial information mediates the impact that access to this information has on memory and metamemory.

2.2.1 Method

2.2.1.1 Participants

This was an online experiment and all potential participants were explicitly instructed to complete the study on a computer (as compared to a phone or a tablet). They were also asked to confirm what device they were using. If a potential participant did not indicate he was on a computer, he was still allowed to complete the task (so as to prevent false reporting) but his data was not collected. Data was collected only for those who completed the entire study. Altogether, there were 102 native English speakers (62 women; mean age = 26.1, *SD* = 9.1) who completed the full study on their computers (in a location of their choice) and, consequently, whose data was recorded. They were recruited via links to the experiment on (i) the University of Leeds Participant Pool Scheme, (ii) websites advertising online psychology experiments (e.g., Psychological Research on the Net and Call for participants) and (iii) social networking sites (i.e. Twitter, Facebook and Reddit).

Participants were randomly allocated to one of two conditions with half the participants told they would be tested on memory for the target's location (Intentional encoding condition) and half the participants not told this (Incidental encoding condition). Participants were not given any compensation for taking part in the experiment except for Psychology students at the University of Leeds who could claim course credit (the distribution of participants taking part for credit and those that did not was equal between the two experimental groups). Feedback was provided at the end of the experiment in the form of a breakdown of memory performance. The study was granted ethical approval by the School of Psychology Ethics Committee, University of Leeds, UK.

2.2.1.2 Materials

For each participant, the studied items were randomly selected from a list of 628 common, singular English nouns (5-6 letters long) taken from the English Lexicon Project (minimum log Hyperspace Analogue to Language frequency 8.02; Balota et al., 2007). This meant each participant was exposed to a unique set of cue-target pairs. This same list of words was also used in Experiment 2 and Experiment 3.

2.2.1.3 Procedure

In a self-paced Study Phase, participants first learned 32 individually presented cue-target pairs. While the cue always appeared in the centre of the screen in red font, the target appeared in one of four locations diagonally from it in black font (see Figure 2.1). The font colours differed so as to make it obvious which was the cue and which was the target (this was explained in the instructions). Each target location was occupied an equal number of times (eight in total). This was followed by a Judgment Phase where, on presentation of each studied cue (this time in black font), participants were asked to (i) indicate in which of the four locations its associated target appeared on screen at study, (ii) indicate a retrospective level of confidence in the correctness of their response (0-20-40-60-80-100%) and (iii) give a JOL prediction indicating their confidence that they would recognize the target on a recognition test (0-20-40-60-80-100%). Lastly, participants completed a forced-choice recognition test in which, for each of the 32 studied cues, they chose the correct target from two options (the distractor was a target of another studied pair to control for baseline familiarity). For each task, the response options were presented as buttons on screen and the participants responded by clicking on the appropriate button with their mouse. There were no time limits imposed on responding.

While half of the participants were informed of the full test procedure before starting the experiment (Intentional encoding), half of the participants were not told they would be asked to remember the location of the studied target (Incidental encoding). Within conditions we also counterbalanced the order of judgments so that half the participants completed the JOL before the target location identification task whereas the other half completed the tasks in the order described above.

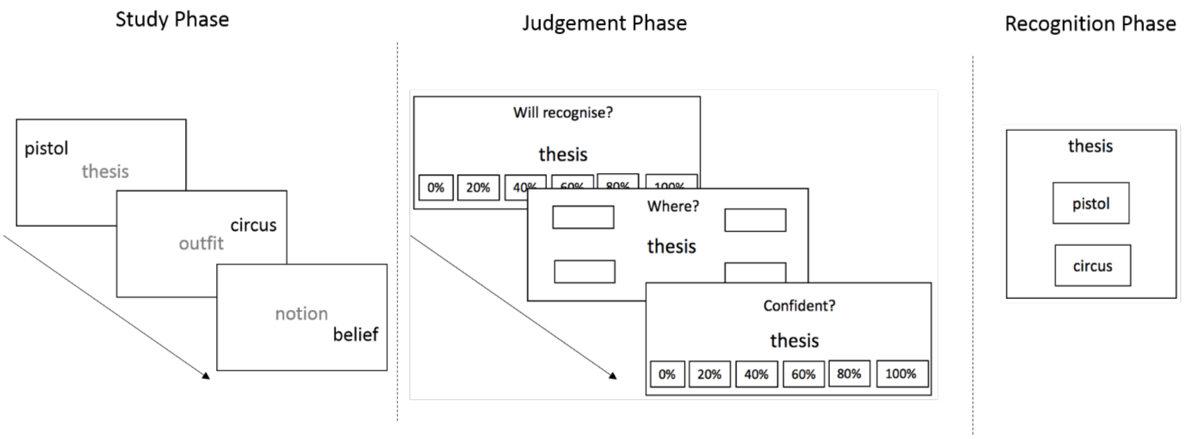

Figure 2.1: Schematic of Experiment 1 procedure. *Cues in the Study Phase were presented in red font (depicted in grey here).*

2.2.2 Results

In this and all subsequent experiments, we first report memory performance in the spatial judgment and final recognition tasks, along with the relationship between retrieval in these tasks. We then apply regression analysis to examine the factors that significantly predict JOL magnitude.

2.2.2.1 Memory performance

Firstly, we assessed whether participants in the two encoding conditions differed in their memory for both the target identity (measured as percentage of correctly recognized targets out of the total 32 trials) and target location (percentage of targets whose location was correctly remembered). An independent samples *t*-test revealed that participants correctly recognized more targets in the Incidental encoding condition (when they were *not* told they would also be tested on their memory for target location; *M* = 87.9%, *SD* = 12.4) as compared to when they explicitly attempted to remember both pieces of information (*M* = 78.3%, *SD* = 16.7), *t*(92.26) = 3.30, *p* < .001, *d* = 0.66. In contrast, participants were not reliably more accurate at remembering the target location correctly when they were told their memory for that information would be tested (*M* = 47.4%, *SD* = 17.0) as compared to when they were not (*M* = 45.0%, *SD* = 19.1), *t* < 1.

2.2.2.2 Relating memory for target identity and target location

In line with the first aim of the study, we investigated whether the ability to remember the location of the target at study was linked to the ability to recognize it on the recognition test. A 2 (condition: Intentional, Incidental) x 2 (location memory accuracy: correct, incorrect) mixed ANOVA was conducted on accuracy of target recognition (percentage of correctly recognized targets; Figure 2.2). The results showed that participants recognized a higher percentage of targets for which they earlier correctly remembered their location, *F*(1, 100) = 23.34, *p* < .001, $\eta_p^2 = .19$. There was an effect of condition, corresponding to improved target recognition accuracy in the Intentional encoding condition, *F*(1, 100) = 10.53, *p* < .01, $\eta_p^2 = .10$. There was no interaction, *F* < 1.

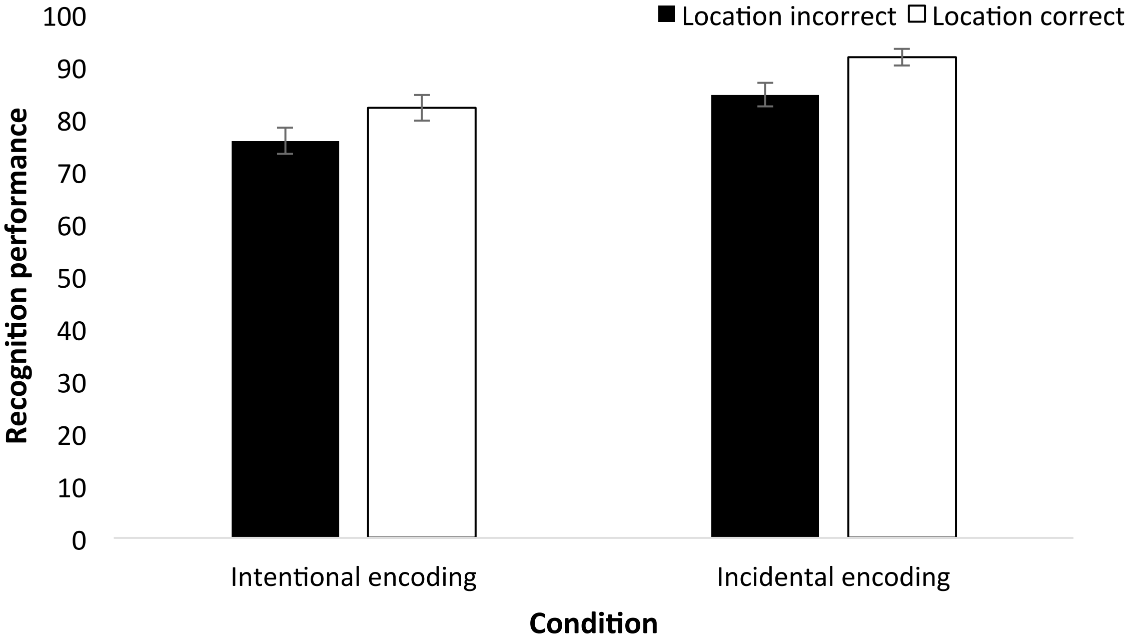

Figure 2.2: The percentage of items correctly recognized as a function of condition and whether their original location was accurately remembered in Experiment 1.

2.2.2.3 Accuracy of spatial memory confidence judgments and JOLs

Participants gave a JOL predicting whether they would recognize the target paired with the presented cue. They also gave retrospective confidence judgments indicating whether they thought they remembered the target's location accurately. We compared the accuracy of both judgments between the two encoding conditions (Incidental, Intentional). The mean AUC results are reported in Table 2.1. The AUC scores for JOLs were the same across the two encoding conditions, $t(100) = 1.71$, $p = .090$, $d = 0.34$. Similarly, AUC scores for the confidence for location judgments were the same across conditions, $t < 1$. A one sample *t*-test showed that all AUC values were above chance (.5) at $p < .001$.

Table 2.1: Mean AUC for JOLs and confidence judgments for location by condition in Experiment 1. *Standard deviations appear in parentheses.*

Encoding condition	JOL for target	Confidence for location
Intentional	.679 (*.196*)	.680 (*.168*)
Incidental	.748 (*.213*)	.700 (*.134*)

2.2.2.4 JOL predictors

It is expected that JOLs accurately track accuracy of target recognition, with higher JOLs expressed for correctly recognized targets. The second key aim of the study was to examine whether JOLs would also increase with having correctly remembered where the target was presented at study and with retrospective confidence in having retrieved this information accurately. To this end a regression analysis was conducted for each condition with JOL magnitude as the outcome measure. Target recognition accuracy (recognized, not recognized), location memory accuracy (correct, incorrect) and retrospective confidence in location memory (0-100%) were the predictors. This analysis was done on a trial-by-trial basis, with a separate regression computed for each participant following a method proposed by Lorch and Myers (1990; see also Allen & Hulme, 2006; Metcalfe et al., 2013). The resulting beta values for each predictor were then analysed using a one-sample *t*-test to determine whether they were significantly different from 0. This enabled a trial-by-trial examination of how a participant's target location memory accuracy, target location memory confidence and their eventual target recognition accuracy, each predicts the JOL they produce, while controlling for inter-participant variability.

Results for the three predictors of interest are presented in Table 2.2. Target recognition accuracy is a predictor of JOL magnitude as would be expected. More importantly, participants' confidence in having remembered the target's location (i.e. perceived access) was also a significant predictor of JOLs. In addition, whether the spatial information was remembered accurately (i.e. target location accuracy) also played a contributing factor to JOL magnitude.

Table 2.2: Mean beta values for each variable included in the within-subject regression analyses of JOL magnitude in Experiment 1 by condition along with one-sample *t*-test results.

Condition	Factor	*β*	*SE β*	*t*	*df*	*p*
Incidental Encoding	Target recognition accuracy	.051	.022	2.35	42	.023
	Location memory accuracy	.049	.020	2.52	50	.015
	Location memory confidence	.621	.040	15.34	49	<.001
Intentional Encoding	Target recognition accuracy	.053	.024	2.20	34	.034
	Location memory accuracy	.094	.025	3.72	49	.001
	Location memory confidence	.574	.044	12.81	48	<.001

2.2.3 Summary

Starting with an examination of memory performance, spatial information was related to recognition accuracy for the target, indicating that access to spatial context information is related to access to item identity. In other words, when presented with the cue, participants can remember both where the associated target was presented as well as what it was, and ability to retrieve spatial information implies ability to retrieve target identity. It is possible that participants were more likely to remember any information about items that were overall better encoded during study. Nevertheless, the observation that this was true for both encoding conditions shows that it does not hinge on intentional encoding of item-spatial information, thus suggesting that this verbal-spatial binding might be relatively automatic in nature. This extends findings from studies on object-spatial binding (e.g., Starns & Hicks, 2008) to verbal material and further lends support to the idea that spatial information is encoded fairly automatically (Köhler et al., 2001).

We also saw a decrease in recognition memory performance in the Intentional as compared to the Incidental condition. In other words, when participants were trying to actively remember both item and spatial information, their item memory suffered. One possible explanation is that being told to encode both target identity and target position might have changed participants' encoding strategies relative to when they were told to only encode target identity. Thus, the Intentional encoding condition may have biased participants toward less effective strategies, focusing on shallow, perceptual features rather than on deeply processing the cue-target content and relationship, which can negatively impact memory performance (Craik & Lockhart, 1972; Thomas et al., 2012).

Turning to metamemory judgments, the results showed that JOLs are related to access to information about the target's position at study. The beta values obtained from the regression analyses showed that participants' confidence in the accuracy of their target location identification was the biggest predictor of their JOLs from the variables considered in the analysis. This result extends previous findings in the metamemory literature by demonstrating that spatial information about where a recently learned target was located, a type of partial information about the target, could impact JOL magnitude. The results also showed that the accuracy of the accessed spatial information relates to JOL predictions. Thomas et al. (2012) found that metamemory judgments do not increase with accuracy of accessed perceptual features (font colour), leading them to suggest that the accessed feature needs to be conceptual (or otherwise tied to the meaning representation of the target item) for accuracy of its access to inform metamemory judgments. Here we extend these findings by showing that JOLs for verbal material increase with accuracy of access to features that fall into the visuospatial domain (namely spatial access) with higher JOLs given to items whose location was correctly identified. Notably, this spatial access was related to item access. While Thomas et al. (2012) did not report the full descriptive data, they did suggest in the discussion that in their study accuracy of access to conceptual attributes also related to recognition performance whereas this was not the case for accuracy of access to their perceptual attributes. Altogether, this would imply that whether accuracy of access relates to metamemory judgments could be dependent on whether that feature access is related to item memory and not on whether it is conceptual in nature (or otherwise tied to the meaning of the target).

The next question of interest was whether there are boundary conditions on the so far observed effects. Köhler et al. (2001) have suggested that not all spatial information might be equivalent, with retrieval of absolute information (where exactly an item was presented) harder than retrieval of relative spatial information (i.e. where in relation to other items it was presented). Consequently, one can distinguish between different types of spatial information, with a possible hierarchy in the ease of spatial access. This in turn suggests that how access to spatial information relates to item memory and JOL magnitude might be dependent on the type of spatial information tested. In Experiment 1, participants could complete the spatial memory test by remembering both where the target appeared exactly on screen (absolute location) and where it appeared in relation to the cue (relative location). In Experiment 2, we investigated whether the pattern of results would hold if test of spatial information no longer indicated precisely where the target appeared on screen but only captured its relative position to the cue. We expected that relative-only spatial information would be harder to access than spatial information that is both relative and absolute.

2.3 Experiment 2

Experiment 2 implemented the same methodology as Experiment 1 with the exception that the location of the cue on screen varied between study trials. More specifically, the cue could appear in one of 16 possible locations on screen. As in Experiment 1, the target appeared in one of four diagonal locations from the cue. At test, the cue was only presented in the centre of the screen, and participants were asked to remember the position of the target on screen in relation to the cue. This meant that the target's location at recall did not correspond to the target's exact (absolute) location on screen at study. This allowed us to examine whether access to spatial information about the target would still be considered meaningful information in JOLs, and relate to target memory, when the assessed location no longer indicated where the target actually appeared at study.

2.3.1 Method

2.3.1.1 Participants

Participants (104 in total, 68 women; 1 undisclosed gender; mean age = 24.6, *SD* = 8.0) were randomly assigned to one of two instructions conditions (52 participants in each). As in Experiment 1, participants completed the experiment online on their computers.

2.3.1.2 Procedure

The general procedure was the same as in Experiment 1. The key change was at study; whereas in Experiment 1 the cue was always presented in the centre of the screen, in Experiment 2 the screen was figuratively divided into a 4x4 grid and the cue could appear in any one of the 16 locations within that grid (see Figure 2.3). The target again appeared in one of four diagonal locations from the cue (top-right, top-left, bottom-right, bottom-left). For example, the cue could appear in the top right corner of the screen in which case the target would have been presented immediately below the cue, diagonally from it to the left. Each diagonal relationship between the cue and the target was shown eight times.

Again, the cue was presented in red font and the target in black font to make it clear which was which. The Judgment Phase was the same as in Experiment 1 with the cue appearing in the centre of the screen in black and participants asked to pick one of the four locations diagonally from the cue at which the target was positioned in relation to it at study. The positioning of the cue and the target at test did not map onto any of the actual cue or target locations at study, which meant the location identification judgments in this experiment were purely about the relation between the two items. Again, participants also provided a confidence judgment about their ability to remember the spatial relationship correctly and a

JOL predicting their ability to recognize the target. The order of these judgments was counterbalanced across participants. The final part of the experiment was a Memory Phase where participants had to choose the corresponding target to each cue from two options. As in Experiment 1, half of the participants were informed they would be asked to remember the spatial cue-target relationship (Intentional encoding condition) and half of the participants were not told this (Incidental encoding condition).

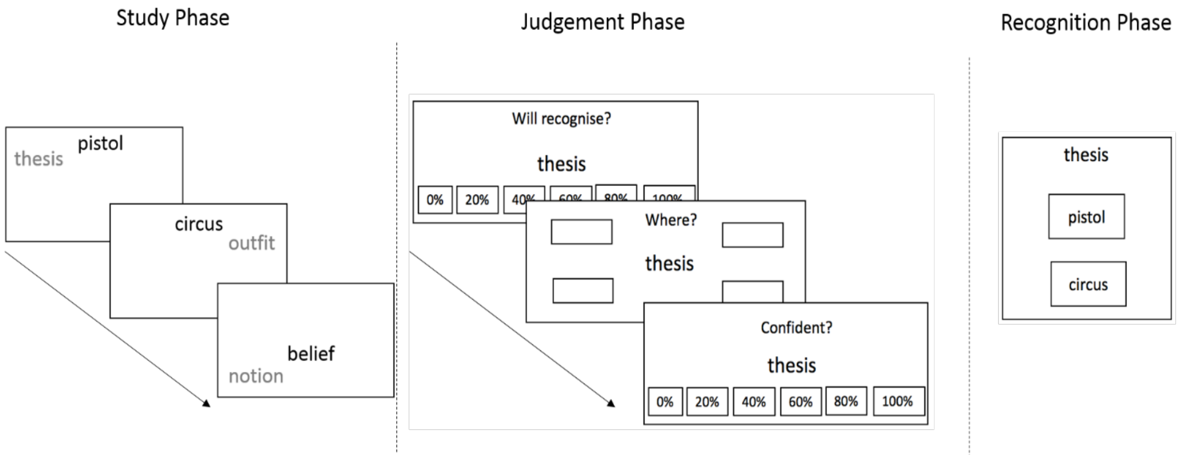

Figure 2.3: Schematic of Experiment 2 procedure. *The cue in the Study Phase is here presented in grey whereas in the experiment it was presented in red.*

2.3.2 Results

2.3.2.1 Memory performance

Firstly, we investigated whether participants differed in their target identity and target location memory performance across the two encoding conditions. An independent samples *t*-test showed that participants recognized the same percentage of targets in the Intentional encoding condition (*M* = 79.8%, *SD* = 16.2) as in the Incidental encoding condition (*M* = 81.1%, *SD* = 16.7), *t* < 1. The percentage of targets whose relative-location was accurately remembered between the Intentional encoding (*M* = 29.8%, *SD* = 11.9) and Incidental encoding (*M* = 33.8%, *SD* = 11.3) conditions was likewise not different, *t*(102) = 1.79, *p* = .076, *d* = 0.35. The performance on the spatial memory task was lower than in Experiment 1 but a one sample *t*-test confirmed that the performance was above chance (25%) in the Intentional encoding, *t*(51) = 2.87, *p* = .006, and the Incidental encoding, *t*(51) = 5.63, *p* < .001, condition.

2.3.2.2 Relating memory for target identity and target location

Secondly, we analysed whether if participants had access to the relative cue-target location information (indexed by spatial memory accuracy) they were also more likely to have access to the target itself (indexed by target recognition accuracy). A 2 (condition: Intentional, Incidental) x 2 (location memory accuracy: correct, incorrect) mixed ANOVA was used to analyse target recognition accuracy (percentage of targets correctly recognized; Figure 2.4). There was no effect of location memory accuracy *F*(1, 102) = 1.65, *p* = .202, $\eta_p^2 = .02$, condition, *F* < 1, or the interaction, *F* < 1. These findings indicate that being able to recognize the target did not imply one also remembered where it was presented at study in relation to the cue, and vice-versa.

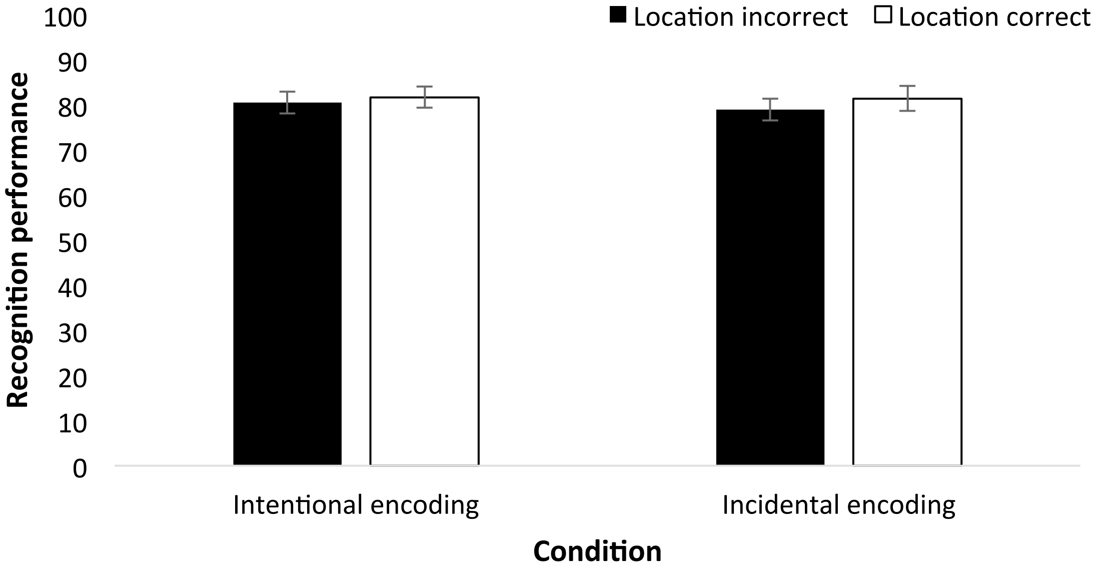

Figure 2.4: Percentage of items correctly recognized as a function of condition and whether their original location was accurately remembered in Experiment 2.

2.3.2.3 Accuracy of spatial memory confidence judgments and JOLs

To analyse accuracy of both judgments, we again compared AUC between the two encoding conditions (see Table 2.3). JOL accuracy was the same between conditions, $t < 1$, as was the confidence for location accuracy, $t(102) = 1.83$, $p = .071$, $d = 0.36$. All the reported AUC scores were above chance (.5) as determined by a one-sample t-test (all p-values < .05).

Table 2.3: Mean AUC for JOLs and confidence judgments for location by condition in Experiment 2. *Standard deviations appear in parentheses.*

Encoding condition	JOL for target	Confidence for location
Intentional	.708 (*.185*)	.584 (*.123*)
Incidental	.680 (*.193*)	.539 (*.123*)

2.3.2.4 JOL predictors

The next question of interest remained whether JOLs would be related to target location memory accuracy and retrospective confidence in having correctly remembered the (relative) target position in addition to target recognition accuracy. As in Experiment 1, trial-level, within-participant regression analyses were used to examine the relative contributions of target recognition accuracy (recognized, not recognized), location memory accuracy (correct, incorrect) and location memory confidence (0-100%) to JOL magnitude. The results (see Table 2.4) again showed that location memory confidence was a significant predictor of JOLs as was target recognition accuracy. This time location memory accuracy did not predict JOL magnitude in either of the conditions.

Table 2.4*:* Mean beta values for each variable included in the within-subject regression analyses of JOL magnitude in Experiment 2 by condition along with one-sample *t*-test results.

Condition	Factor	*β*	*SE β*	*t*	*df*	*p*
Incidental Encoding	Target recognition accuracy	.086	.020	4.34	42	<.001
	Location memory accuracy	.025	.020	1.23	50	.225
	Location memory confidence	.584	.032	18.16	50	<.001
Intentional Encoding	Target recognition accuracy	.053	.021	2.55	41	.015
	Location memory accuracy	-.017	.023	-0.74	49	.465
	Location memory confidence	.591	.044	13.45	48	<.001

2.3.3 Summary

In Experiment 2 we observed that memory for relative spatial information was not related to memory for the target itself as assessed by performance on the recognition task. This builds on results of Experiment 1 (where both relative and absolute spatial information was accessible), and shows that while probability of spatial and identity access can be related this is not always the case, depending on the type of spatial information tested.

This time there were no effects of intentionality of encoding condition on recognition memory performance (which was lower in the Intentional encoding condition of Experiment 1). In neither experiment did we instruct participants on what strategies to use to encode the cue-target pairs. A possible explanation for the results of Experiment 1 was that participants adopted different encoding strategies between the encoding conditions. This would imply that participants who took part in Experiment 2, might have adopted similar encoding strategies independent of condition, leading to equivalent recognition performance.

In regards to metamemory judgments, Experiment 2 again demonstrated that perceived access to relative spatial information (measured here by participants' confidence in the accuracy of their location judgements) relates to JOL magnitude. However, in Experiment 2 we did not observe any effect of accurately remembering where the target appeared in relation to the cue on JOL magnitude once other factors were taken into account. This is in contrast to Experiment 1 where we did observe this effect. However, in Experiment 1 spatial access was related to target recognition memory whereas in Experiment 2 it was not. This gives further support to the idea that the determining factor in whether accuracy of access relates to JOLs is not dependent on the type of information accessed but solely on whether it is related to item access.

In summary, the results of Experiment 2 helped to further clarify the relationship between metamemory and memory. When only relative spatial target information was available, access to this information was no longer related to item memory and correspondingly, accuracy of this access did not relate to JOL magnitude. Experiment 2 differed from Experiment 1 in that (i) retrieval of absolute target location was not supported and (ii) the spatial information (being relative only) was harder to access (as indicated by spatial memory performance). It is not clear which of the two factors could drive the differences in the observed results. The aim of the final experiment was to understand how the observed pattern of results compared to a condition in which cue location again varied in an uninformative manner but only absolute target location (i.e. exact position on screen) was assessed. To allow for direct comparisons, we not only tested this new condition but also replicated results of Experiment 1 and 2.

2.4 Experiment 3

Experiment 3 aimed to replicate the outcomes of the first two experiments, and extend to a condition where only the target's absolute location was assessed. If support for absolute spatial access is a key differentiation between Experiment 1 and 2 then access to absolute-only spatial information should be related to item access and accuracy of this access should impact JOL magnitude. A possible explanation for this could be that absolute spatial location of the target is more closely linked to the target representation than information on how it *relates* to other items such as the cue. Alternatively, as absolute information may be more difficult to access than relative information (Köhler et al., 2001), the results might be similar to Experiment 2, with no link between (absolute-only) spatial access and item access. Regardless, we expected to see the same effect of perceived spatial access on JOL magnitude as in previous experiments in all conditions.

The instructions manipulation was removed as it was a secondary focus in Experiments 1 and 2, and did not have any major effects on the metamemory outcomes. Given that target recognition performance seems to have somewhat suffered in the Intentional encoding conditions (in Experiment 1), we focused only on the Incidental encoding condition. Finally, we reduced the possible target locations from four to three, and increased the number of choices at the recognition test from two to three, in an attempt to reduce the difference in difficulty levels between these memory tasks.

2.4.1 Method

2.4.1.1 Participants

Overall, 153 participants (101 women, 1 undisclosed gender; mean age = 25.5, *SD* = 7.3) took part in the study. Of these 51 participants were assigned to the Full reinstatement condition (same as Experiment 1), 52 were assigned to the Relative-only condition (same as Experiment 2) and 50 were assigned to the Absolute-only condition. As in previous experiments, participants completed the experiment online on their computers.

2.4.1.2 Procedure

The procedure was an adaptation of the first two experiments with the key manipulation at study. Participants first studied a series of 33 cue-target pairs (an increase from Experiment 1 and 2 to allow for equal number of presentations for each of the spatial locations). In the condition that replicated Experiment 1 (here termed the Full reinstatement condition), the cue always appeared in the centre of the screen whereas the target appeared in one of three diagonal locations from the cue (the three locations were randomly chosen out of the four possibilities for each participant). In the condition that replicated Experiment 2 (here termed the Relative-only condition) the cue appeared in one of 16 possible locations on the screen (within a 4x4 grid) and the target in one of three possible diagonal locations from it. In the new, Absolute-only condition, while the target again appeared in one of three diagonal locations from the centre indicated by a fixation point (same as in the Full reinstatement condition), this time the cue did not appear in the centre but in any other diagonal location from the target (see Figure 2.5). Again, the cue was always presented in red font and the target in black font to distinguish the two.

Following study, participants were presented with each of the studied cues in the centre of the screen and were asked to indicate where the target appeared at study. In the Absolute-only condition the options corresponded to actual target locations. In the Relative-only condition the options corresponded to where the target appeared in relation to the cue but did not represent where the target was located on the screen. In the Full reinstatement condition, the options corresponded to both the absolute-only and relative-only target position at study. Participants also indicated their confidence in their spatial memory and gave JOLs concerning whether they thought they would recognize the target in the recognition test. The last part of the experiment was a recognition test where for each cue participants were asked to select the target from three options (all distractors were targets from the study).

We removed the instructions manipulation so that no participants were told to remember the spatial information. As in previous experiments, the order of the judgments was counterbalanced across participants.

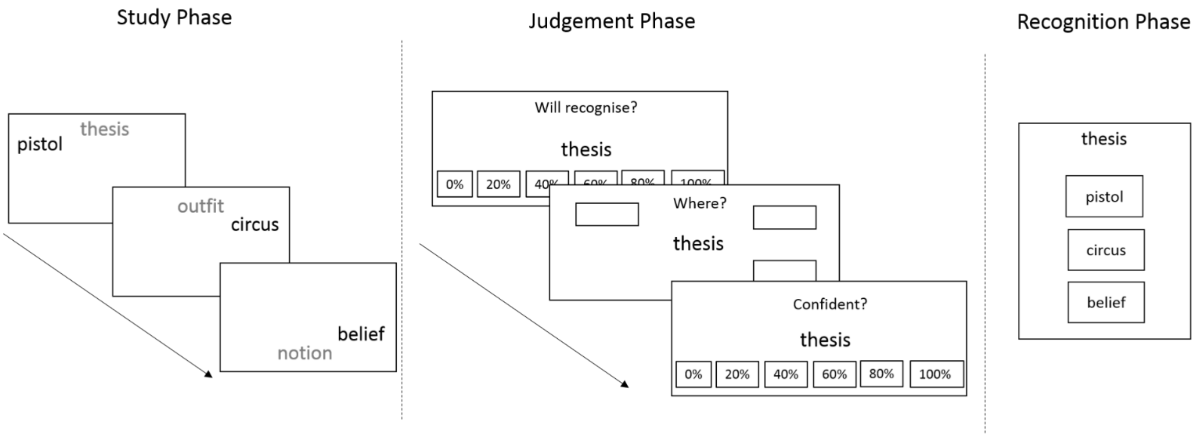

Figure 2.5: Schematic of Experiment 3 procedure. *The Study Phase schematic is for the Absolute-only condition (target position corresponds exactly to presented options at test). The cue in the Study Phase is here presented in grey whereas in the experiment it was presented in red.*

2.4.2 Results

2.4.2.1 Memory performance

Firstly, we compared overall memory performance between the three conditions. An independent samples ANOVA showed that percentage of correctly recognized targets was the same between the Absolute-only (*M* = 78.7%, *SD* = 18.1), the Relative-only (*M* = 80.1%, *SD* = 22.4) and the Full reinstatement (*M* = 81.2%, *SD* = 20.5) conditions, *F* < 1. However, participants performed better in identifying the targets' combined absolute and relative location as seen in the Full reinstatement condition (*M* = 51.3%, *SD* = 16.2) as compared to only the relative target location (*M* = 40.5%, *SD* = 14.1), and the absolute-only target location (*M* = 34.3%, *SD* = 8.7), *F*(2, 150) = 20.95, *p* < .001, $\eta_p^2 = .22$. Pairwise comparisons using the Bonferroni correction indicated that the latter two groups differed from each other only marginally (*p* = .063) whereas both were lower than the performance in the Full reinstatement condition (*p* < .001). However, whereas performance in the Relative-only condition was above chance (33%), *t*(52) = 3.85, *p* < .001, performance in the Absolute-only condition was not, *t*(49) = 1.06, *p* = .293.

2.4.2.2 Relating memory for target identity and target location

As in the previous experiments, we analysed target recognition performance by whether that target's location was correctly remembered (Figure 2.6). A 3 (condition: Full-reinstatement, Absolute-only, Relative-only) x 2 (location memory accuracy: correct, incorrect) mixed ANOVA analysing recognition performance revealed marginal effect of location memory accuracy, *F*(1, 150) = 3.85, *p* = .051, $\eta_p^2 = .03$. There was no effect of condition, *F* < 1, but there was an interaction between the two factors *F*(2, 150) = 7.22, *p* = .001, $\eta_p^2 = .09$. For the Full-reinstatement condition, access to spatial information was related to access to the actual target as indicated by recognition performance, $t(50) = 3.82$, $p < .001$, $d = 0.56$, whereas in the Relative-only, $t < 1$, and the Absolute-only, $t < 1$, conditions this was not the case. These analyses confirm and extend the findings of the first two experiments.

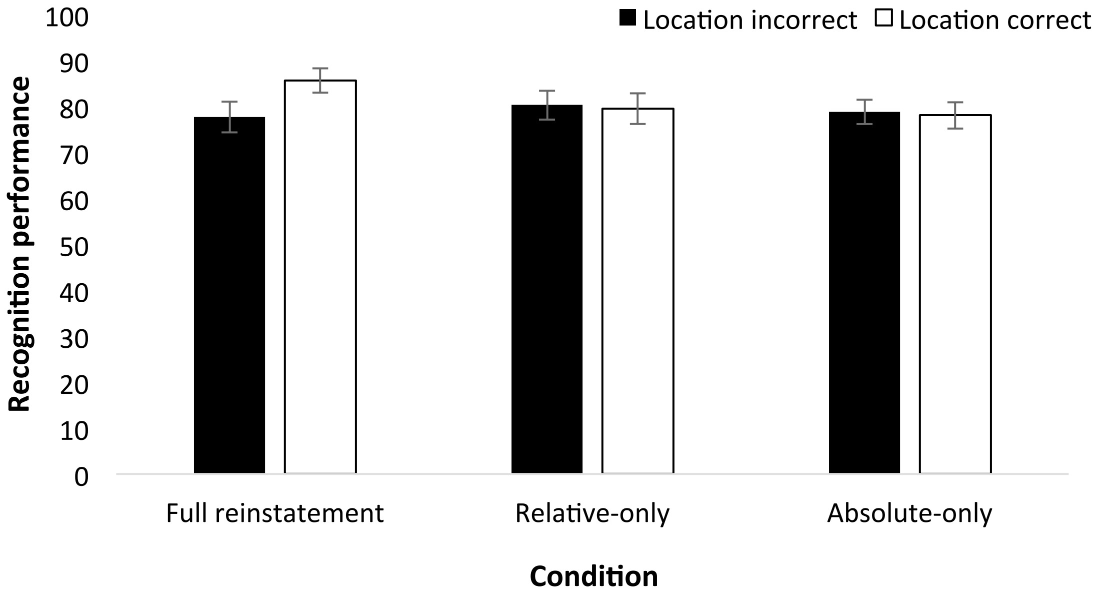

Figure 2.6: The percentage of items correctly recognized as a function of condition and whether their original location was accurately remembered in Experiment 3.

2.4.2.3 Accuracy of spatial memory confidence judgments and JOLs

Mean AUC scores assessing accuracy of JOLs and confidence judgments are reported, per condition, in Table 2.5. A one-way between subjects ANOVA showed that there was no difference between conditions in JOL accuracy, $F < 1$. There was however a significant difference between conditions in the accuracy of the location confidence judgments, $F(2, 150) = 8.73$, $p < .001$, $\eta^2 = .95$. Post-hoc tests using the Bonferroni correction showed that participants in the Absolute-only condition were significantly worse than participants in the Full Reinstatement ($p < .001$) and the Relative-only ($p = .004$) conditions. There were no further differences.

Further, a one-sample *t*-test was employed to determine whether all AUC values were significantly above chance performance (0.5). JOLs across all conditions were accurate above chance (all *ps* < .001). In contrast, the confidence location AUCs were above chance for the Full Reinstatement (*p* < .001) and Relative-only (*p* < .001) conditions but not in the Absolute-only (*p* = .062) condition.

Table 2.5: Mean AUC for JOLs and confidence judgments for location by condition in Experiment 3. *Standard deviations appear in parentheses.*

Encoding condition	JOL for target	Confidence for location
Full Reinstatement	.717 (*.190*)	.636 (*.139*)
Relative-only	.746 (*.181*)	.619 (*.145*)
Absolute-only	.733 (*.182*)	.596 (*.117*)

2.4.2.4 JOL predictors

As before, we analysed the combined contributions of the variables of interest to JOL magnitude. In other words, we investigated whether, for each condition, it could be said that the JOL was related to a number of sources of information including target recognition accuracy (which it aims to predict), target location memory accuracy (which in some conditions is related to target memory) and retrospective location memory confidence for the targets (i.e. perceived access). The results of the within-participant, trial-level regression analyses are presented in Table 2.6. The Full reinstatement condition results mirror those of Experiment 1 as all three variables were shown to significantly predict JOL magnitude. The Relative-only condition similarly mirrors results of Experiment 2 as we again observe the effect of location memory accuracy on JOLs disappearing. Lastly, the Absolute-only condition shows the same pattern of results as the Relative-only condition with target recognition accuracy and location identification confidence being the only significant predictors of JOLs from the variables included in the analysis.

Table 2.6*:* Mean beta values for each variable included in the within-subject regression analyses of JOL magnitude in Experiment 3 by condition along with one-sample t-test results.

Condition	Factor	*β*	*SE β*	*t*	*df*	*p*
Full reinstatement	Target recognition accuracy	.080	.018	4.54	37	<.001
	Location memory accuracy	.099	.021	4.67	50	<.001
	Location memory confidence	.582	.035	16.86	50	<.001
Relative-only	Target recognition accuracy	.177	.019	9.07	47	<.001
	Location memory accuracy	.020	.017	1.18	51	.244
	Location memory confidence	.482	.039	12.25	51	<.001
Absolute-only	Target recognition accuracy	.159	.025	6.42	39	<.001
	Location memory accuracy	.002	.023	0.10	48	.925
	Location memory confidence	.535	.036	14.84	48	<.001

2.4.3 Summary

The results of Experiment 3 confirm and extend those of Experiments 1 and 2. It appears that while memory for combined absolute and relative spatial information is related to recognition memory for the target, the relative-only and absolute-only cue-target spatial memory is not. Memory accuracy for spatial information only related to JOL magnitude in the Full reinstatement condition where we observed this dependency between spatial and item access.

The relationship between location memory confidence and JOL magnitude was true for all conditions and shows that confidence in accessed partial information (even when this information is not related to memory performance) can increase with confidence that the target will be recognised.

2.5 Discussion

This study investigated whether and how access to spatial location of the target at study relates to recognition accuracy for the target and to JOL magnitude. Given the relevance of spatial context to how we characterize episodic memory (Burgess et al., 2001; Maguire & Mullally, 2013; Robin et al., 2016; Tulving, 1985), it is important to understand whether remembering where the learned item was located at study relates to confidence that it will be retrieved later and whether it relates to actual memory for the item. We examined whether this was true for both relative and absolute spatial information. Our main findings were as follows:

1. Participants can remember item location, even when this is incidental to the encoding instructions, except for absolute-only location.
2. The accuracy of this item location identification is positively related to the accuracy of a subsequent cue-target recognition task, but only in the Full Reinstatement condition where both absolute and relative spatial information is available.
3. Perceived access to target's position at study (indicated by participants' confidence in their location judgment) significantly predicted JOLs in all conditions, irrespective of actual memory performance for this information. This is true at an individual trial level within each participant when the contribution of actual memory performance (which is being predicted) is taken into account.

4. Accuracy of spatial access relates to JOL magnitude only in instances when it is also related to item access (the Full Reinstatement condition).

These novel findings provide a number of new insights concerning episodic memory and metamemory. We focus first on the memory results and then move onto the metamemory outcomes.

Past research has shown that spatial information is bound to the item while not necessarily being integrated with other source information (Starns & Hicks, 2008) and that this binding occurs at encoding (Uncapher et al., 2006). The present study extends these findings by demonstrating that participants were able to encode where learned information was presented and could retrieve this information even when not initially instructed to attend to it during encoding. Furthermore, those items for which participants were able to successfully retrieve location information were then more likely to themselves be correctly recognised, at least in the case of the Full Reinstatement condition (Experiments 1 and 3). These findings indicate a process of binding between cue and target within a spatial configuration, producing multi-element associative representations or engrams (e.g., Horner & Burgess, 2013; Tulving, 1983) containing information about identity and location. The processes that operate at encoding to initially construct such representations may be relatively automatic in nature (Köhler et al., 2001), as similar outcomes are apparent in both intentional and incidental conditions.

However, the relationship between access to spatial information and target recognition accuracy was mediated by the type of location task that was implemented; the likelihood of remembering where the target was presented was only related to memory for target identity in the Full Reinstatement condition, and not when the spatial information tested was for either the target's absolute or relative spatial information only. This may be related to relative ease of access. We observed that making location judgments when both relative and absolute spatial information was accessible (in the Full Reinstatement condition) was easier than accessing absolute-only or relative-only spatial information in isolation. Indeed, performance for absolute-only spatial information was at chance, suggesting that relative-only spatial information is somewhat easier to access, in line with previous findings (Köhler et al., 2001). A possible explanation for this is that retrieval of absolute location requires access to object identity whereas retrieval of relative item position can be retrieved even when item identity is not available (Köhler et al., 2001). While we cannot directly speak to participants' ability to recall target items as this was only tested via recognition, we can nevertheless conclude that relative-only processing might be favoured to absolute-only processing, and that removing either element reduces spatial memory performance relative to conditions of full reinstatement. This increase in difficulty of access to location seems to minimize any relationship between location memory accuracy and subsequent target memory recognition. Thus, the ability to remember *where* is only related to the ability to remember *what* when the former information is relatively more accessible and both relative and absolute routes to retrieval are made available.

A primary focus of the current work was the exploration of the factors that relate to judgment of learning (JOL) magnitude concerning cue-target recognition performance. The results from all three experiments demonstrated that participants' JOL magnitude, predicting that they would recognize the target on a later memory test, directly increased with their belief that they could retrieve spatial information about the recently learned cue-target word-pairs. These results were demonstrated when the data was analysed on a trial-level in an item-by-item analysis within participants, meaning that observed predictors of JOLs were independent of variability across participants (such as some participants giving overall higher JOLs). This finding held when participants were not instructed to attend to the spatial information at study. Furthermore, this held true when both absolute and relative spatial information was accessible (i.e. the Full Reinstatement condition) as well as when the test was only for the targets' absolute or relative positions and access to the spatial information did not relate to recognition accuracy for the target. Lastly, perceived spatial access was shown to significantly predict JOL magnitude even when the contributions of target recognition performance and spatial memory accuracy to explaining JOL variance were also considered. This finding adds to the growing literature on the type of information that relates to metamemory judgments.

The results of the current study further provide evidence that delayed JOLs can change with heuristics not indicative of future item memory. While this has been demonstrated for immediate JOLs (i.e. judgments made during or immediately after study of each item and before next item is presented; Alban & Kelley, 2013; Koriat & Bjork, 2006; Rhodes & Castel, 2008), it has not yet been demonstrated in delayed JOLs. Overall, these finding are in line with the accessibility view of metamemory (Koriat, 1993) suggesting that the more information one can access at time of judgment (irrespective of accuracy), the more confident one is that they know the target.

Additionally, the study extends the findings of Thomas et al. (2012), which demonstrated that in particular cases, the quality or accuracy of information access can also increase metamemory judgments. They found that the accuracy of accessed conceptual (target category) but not perceptual (font colour) information about the target increased FOKs. The experiments presented here demonstrated that the accessed information does not have to be conceptual but can also fall into the visuospatial domain for its accuracy to affect metamemory judgments. More specifically, we observed that accuracy of access to spatial information about the target increased JOL magnitude but only when the spatial access was related to recognition accuracy for the target (as seen when both absolute and relational spatial information was tested as in Experiment 1 and the Full-reinstatement condition of

Experiment 3). Based on these results we suggest that the determining factor in whether accuracy of access relates to metamemory judgments is how that access relates to item memory.

A caveat to the present findings is that by asking participants about specific characteristics of the learned information (e.g., the target's spatial location), we are making that information particularly salient. It is less clear whether participants assess access to this type of contextual information when making a JOL if not asked about it. This issue is relevant to the majority of metacognitive literature at present (see for example Hertzog et al., 2014) and relates to the fact that we continue to have to rely on explicit report from participants to be able to investigate the variables of interest. This is discussed in more detail and addressed in Chapter 4. In the context of the present study, we tried controlling for this issue by (i) counterbalancing the order of the two judgments across participants and (ii) by manipulating whether participants were instructed to pay attention to the spatial information at encoding. The observed results were clearly independent of the encoding condition manipulation.

Another limitation is the correlational nature of the results. This means that we can only speak to their being a relationship between the factors of interest without being able to make strong causality statements. However, using a regression has allowed us to show that the confidence in spatial memory significantly predicted JOL magnitude even when other variables were taken into account. What is more, the other variables in the analysis (memory for the target and memory for its spatial location) were the type of variables that classic theories of delayed JOLs would argue should be the primary determinants of such judgments as they relate to (or at least appear to relate to) target memory strength. That spatial memory confidence significantly predicted JOL magnitude when these other variables were included in the analysis suggests that access to contextual information can influence metacognitive judgments.

In summary, the results of this study add to the growing literature exploring which type of information access and item features can impact metamemory judgments (Alban & Kelley, 2013; Jersakova et al., 2015; Koriat et al., 2003; Schwartz, Pillot, & Bacon, 2014; Thomas et al., 2012) by demonstrating that access to contextual (spatial) features relates to an increase in delayed JOL magnitude. As such we extended the noncriterial recollection hypothesis developed in the context of FOK research to the delayed JOL paradigm. Further, the study suggests that considering the combined influences of perceived access and accuracy of feature access on metamemory predictions allows for a fuller understanding of how metamemory judgments are constructed. We also specified a possible mechanism that modulates the relationship between access accuracy and metamemory judgments (namely its relationship to item memory). Nevertheless, we also note some of the problems of this type of approach that is relevant to much of the metacognitive literature.

CHAPTER 3

EXPERIMENTAL vs. PRE-EXPERIMENTAL CUE FAMILIARITY

3.1 Introduction

In many metacognitive paradigms, the cue is the only information participants are reliably provided with and the one piece of information that researchers can be confident participants have access to at time of judgment. It is assumed that participants attempt to retrieve the target and related information but unless asked directly about their access, we do not know whether that information was activated when the JOL was made. Correspondingly, the particular role of the cue in metamemory judgments has garnered much attention and Schreiber & Nelson (1998) suggested a broad category of 'cue effectiveness' effects that impact metacognitive judgments. For example, Schreiber and Nelson (1998) found that FOK and POK judgments were sensitive to the number of concepts semantically linked to the test cue, with cues that were linked to smaller sets leading to higher judgments. They suggested that more concepts lead to more competition, which in turn lead to a decrease in metacognitive confidence. Similarly, the likelihood of TOT experiences was found to increase with the amount of information contained in the cue used to elicit the target, even if this increase in information was redundant (e.g., repetitions in general knowledge definitions used to elicit search for the target term; Koriat & Lieblich, 1977). These types of effects highlight the importance of the cue even in judgments made specifically about the accessibility of the target.

Given the nature and prevalence of the cue-target paradigm in the metacognitive literature, the present chapter explores the influence of the cue on memory for the target item and on JOLs. More specifically, we focused on two types of cue familiarity (experimental and pre-experimental) and compared them directly thus speaking to two literatures that have to date developed independently from each other. Across two experiments we manipulated experimental cue familiarity by presenting half of the cues in a pre-study rating task, thus increasing their familiarity at time of study and throughout the task compared to the rest of the cues. We also manipulated pre-experimental familiarity by using either completely novel, pseudo-word cues (Experiment 4a) or real-word cues (Experiment 4b). Altogether this allowed us to examine how different types of cue familiarity relate to memory and metamemory in paired-associate paradigms.

Vast majority of cue effects in metacognitive research have focused on experimental familiarity i.e. manipulating the level of exposure to some cues as compared to others within the experiment (Metcalfe, Schwartz, & Joaquim, 1993; Reder & Ritter, 1992; Reder & Schunn, 1996; Reder, 1987; Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997). Most commonly experimental familiarity is manipulated through a pre-task exposure to some items that later form the cues (e.g., in a seemingly unrelated rating task) or through manipulating duration and frequency of cue presentation during the study phase (Benjamin, 2005; Liu, Su, Xu, & Chan, 2007; Metcalfe et al., 1993; Metcalfe & Finn, 2008b; Reder & Ritter, 1992; Vernon & Usher, 2003). This wealth of research has demonstrated that metacognitive judgments are strongly influenced by experimentally manipulated familiarity with the cue term used to elicit the judgment; the higher the familiarity with the cue, the higher the likelihood of a positive metacognitive response. This is true even in cases where the cue actually requests access to novel target information, achieved, for example through changing one word of an otherwise familiar general knowledge question or a riddle (e.g., Vernon & Usher, 2003).

The effect of cue familiarity has first been demonstrated in speeded judgments i.e. predictions made under a time limit with the aim of capturing the nature of pre-retrieval mechanisms (Reder & Ritter, 1992). Given that the cue is the first and only information presented at judgment it is not surprising that it should play a role in these initial, guiding processes. Since this earliest work however, the role of experimental cue familiarity has been extended to other, non-timed paradigms starting with FOK judgments made for inaccessible items (Metcalfe et al., 1993) but also evidenced in confidence judgments made for just retrieved targets (Chua et al., 2012) and judgments made at learning (Metcalfe and Finn, 2008b). Current models of metacognitive judgments made about memory highlight the interacting roles of cue familiarity and target accessibility (Benjamin, 2005; Koriat & Levy-Sadot, 2001; Metcalfe & Finn, 2008b) and it is accepted that across paradigms, the cue plays an important guiding role in metacognitive judgments, particularly early in the judgment process.

Despite the general similarities across the episodic paradigms, there remains disagreement as to whether experimentally manipulating cue familiarity affects access to the target with some studies observing this effect (e.g., Metcalfe & Finn, 2008b), some studies finding no effect (e.g., Benjamin, 2005) and yet others reporting mixed results where secondary variables (time given to make a judgment) impacted whether an effect of cue familiarity on target accessibility was found (e.g., Liu et al., 2007). In memory research, experimental manipulations of familiarity have primarily investigated how it influences the study of pre-experimentally novel information. For example, Ebbinghaus (1885/1964) demonstrated that speed of learning of new, nonsense syllables improved with repeated experimental exposure to any given item. Most recently, one study has observed that experimental familiarity is also important for building new associations (Reder, Liu, Keinath, & Popov, 2016). More specifically, participants studied cue-target pairs where the cue was a pair of (to them pre-experimentally novel) Chinese characters and the target an English word. Across multiple study-test sessions over a number of weeks in each of which participants studied new character-word associations, participants performed better at associating the words to characters they have previously been exposed to (even when the association they were asked to form was novel to them). In summary, the effects of experimental familiarity on memory are inconclusive.

In contrast, the advantage of pre-experimental familiarity to memory has been demonstrated in working memory (Allen, Havelka, Falcon, Evans, & Darling, 2015; Darling, Allen, Havelka, Campbell, & Rattray, 2012; Ricks & Wiley, 2009) and long-term memory for single items (Rawson & Van Overschelde, 2008; Ricks & Wiley, 2009; Van Overschelde, Rawson, Dunlosky, & Hunt, 2005) as well as for associations such as memory for item source or contextual information (DeWitt, Knight, Hicks, & Ball, 2012; Reder et al., 2013). Some have argued that semantic memory (pre-experimental familiarity) is a prerequisite for and facilitates episodic encoding (Hintzman, 1988; Moscovitch et al., 2005; Tulving & Markowitsch, 1998); a conclusion dating back to research demonstrating that general knowledge schemas of the world influence how well and what information is encoded (Bartlett, 1932). One of the classic examples of how prior knowledge improves encoding of new information comes from studies on domain-specific learning comparisons between experts and novices (e.g., chess masters' memory for location of chess pieces). Across knowledge domains, experts tend to do better for new information learned in that domain than novices (Chase & Simon, 1973; Voss, Vesonder, & Spilich, 1980).[1] Similarly, it is easier to recall episodic encoding details for familiar compared to completely novel proverbs

[1] A related example is the finding that items are better encoded if preceded by or presented alongside congruent information as compared to incongruent information (Bein et al., 2015; Bransford & Johnson, 1972; Staresina, Gray, & Davachi, 2009). In other words, pre-experimentally formed expectations and knowledge schemas can influence encoding.

(Poppenk, Köhler, & Moscovitch, 2010) and context reinstatement aids face recognition of well-known faces of celebrities but not of newly learned faces (Reder et al., 2013). Overall, it has been repeatedly demonstrated that it is easier to encode new information when it can be integrated with pre-existing knowledge.

The effect of pre-experimental familiarity on metacognition is less explored but some studies have similarly addressed the question of how expertise influences metamemory. The earliest study investigating the impact of expertise in general knowledge information domains found no effect on FOK accuracy (Roberts & Rhodes, 1989). This was confirmed in a more recent study which similarly found no effect of self-efficacy beliefs across knowledge domains on FOK accuracy in a semantic, general knowledge task (Marquié & Huet, 2000). Peynircioğlu and Tekcan (2000) tested native Turkish speakers with varied levels of English proficiency on Turkish-English word translations and found that proficiency impacted the magnitude of FOK judgments but not their accuracy. More recently, studies have also started investigating the role of expertise in episodic metamemory. More specifically, novice and expert chess players were asked to study moves in a chess end-game and predict that they would remember the moves in the future (i.e. make an immediate JOL); not only were experts more accurate, novice players' accuracy was not significantly different from 0 (de Bruin, Rikers, & Schmidt, 2007). On the other hand, a study which asked participants to study new facts from a range of knowledge domains in which their level of expertise varied (as indicated by self-report), observed consistent overconfidence in participants' predictions, which was more pronounced for items from well known as compared to less known topics (Shanks & Serra, 2014). As such the effects of expertise and pre-experimental familiarity on metacognition remain unclear and are likely to differ between tasks. Furthermore, there has not been a study to date that has explored the question of pre-experimental familiarity in an episodic delayed JOL paradigm.

In summary, while experimentally manipulating cue familiarity always influences metacognitive judgments, the influence on memory is less clear. In contrast, while pre-experimental familiarity seems to always aid encoding, its effect on metamemory remains undecided. Further, majority of studies have investigated experimental and pre-experimental familiarity separately. An exception is a study by Poppenk et al. (2010) who compared memory for pre-experimentally familiar proverbs to memory for completely novel proverbs. Half of the novel proverbs were shown to participants in a pre-study rating task increasing their experimental familiarity. Source memory was the same for pre-experimentally and experimentally familiar proverbs, both of which were remembered better than the novel items. The authors concluded that pre-experimental and experimental familiarity could provide a similar kind of encoding advantage. However, except for their study, no one to date has compared the two directly, especially not in an associative episodic memory paradigm. As such the generalizability of this conclusion remains in question. Further, no one has explored the two types of familiarity and their combined influence on metacognitive judgments. The present chapter aims to fill this gap by comparing the effects of experimental and pre-experimental familiarity on memory and metamemory.

A notable aspect of the metacognitive studies looking at the effects of cue familiarity is the idea of change in the influence on JOLs across time. Vernon and Usher (2003) demonstrated that metacognitive judgments can undergo a number of developments and transformations in the weighing of accessible evidence (e.g., cue familiarity) across periods of up to 12 seconds. This demonstrated that metacognitive judgments are dynamic in nature and comparing judgments across a number of time windows can ensure some of the subtleties of the evidence evaluation processes underlying predictions such as JOLs are not missed. Correspondingly, in addition to manipulating experimental and pre-experimental familiarity, this study manipulated the amount of time given to participants to make their JOLs. More specifically, drawing on Koriat & Levy-Sadot (2001) and Vernon and Usher (2003), we employed a speeded condition where participants had to give their JOL within two seconds of cue presentation and a delayed condition where participants had to first wait two seconds before giving their response (again within two seconds i.e. in a 2000-4000ms window from the cue appearing on screen. The speeded time window is slower than that employed in the most similar paradigms to the current study (Benjamin, 2005; Metcalfe & Finn, 2008b) which have focused primarily on capturing initial, pre-retrieval mechanisms in paradigms where participants are not given enough time to do more than register the cue. We were instead interested in observing whether the changes in cue familiarity effects would be more temporally extended.

In summary, the present chapter revisits and extends some classic findings in the metacognitive literature through exploring the effects of experimental and pre-experimental cue familiarity on JOLs and associative memory. Experiment 1 employed pseudo-word cues (i.e. pre-experimentally novel cues) and these were contrasted with (pre-experimentally familiar) real-word cues in Experiment 2. Both experiments also manipulated experimental familiarity through a pre-study rating task, which exposed participants to half of the to-be-studied cues. Throughout the chapter, cue familiarity (high vs. low) refers specifically to the experimental cue familiarity manipulation that was common to both Experiment 4a and 4b. In contrast, the pre-experimental familiarity, between-experiment manipulation is referred to as the cue type (real word vs. pseudo-word cues) manipulation. We also investigated whether these influences would change when participants had to respond within two seconds of cue-presentation at judgment (speeded condition) as compared to when they first had to wait two seconds before responding within a two second window (i.e. between 2000-4000ms after cue presentation (delayed condition). Due to the speeded nature of the paradigm, this time we employed binary (*yes/no*) JOL predictions rather than confidence JOLs. We were interested in how the two familiarity manipulations affected (i) memory for the target, (ii) the likelihood of participants giving a positive (*yes*) JOL prediction, (iii) the response criterion adopted by participants, (iv) their metacognitive accuracy and (v) whether these effects would differ between the two time windows.

Given previous memory research we predicted that both experimental and pre-experimental familiarity would be linked to an improvement in memory performance for the target but that these effects might interact. More specifically, we expected that memory would be better for the real-word cue-target pairs as compared to pseudo-word cues paired with real-word targets. We also anticipated that manipulating experimental familiarity of the pseudo-word cues would improve memory for the targets. Further, we predicted that experimental familiarity would increase positive JOL predictions in both cue type conditions. We were interested in observing whether this would be reflected in a response criterion shift, suggesting a change in response strategies. Lastly, we expected the cue familiarity effects to be strongest in the speeded conditions and less pronounced in the delayed conditions and to see this difference especially for the pseudo-word cues.

3.2 Method

3.2.1 Participants

All participants were students at Université de Bourgogne; 34 (7 men, mean age = 20.2, *SD* = 2.1) participated in Experiment 4a and 32 (3 men, mean age = 19.7, *SD* = 2.0) participated in Experiment 4b. They were all native French speakers, took part for course credit and were only allowed to participate in one of the experiments. The study was granted ethical approval by the University of Leeds ethics review board.

3.2.2 Materials

For Experiment 4a, a list of 122 French pseudo-words was created. The pseudo-words were all six letters long and consisted of two or three syllables. We used pseudo-words rather than rare words because we wanted to use items that were completely novel to participants, items that they could not have had any previous experience with. The list of pseudo-words was generated using the trigram algorithm available through the Lexique Toolbox (New & Pallier, 2001b). The algorithm takes real words provided by the user and creates a list of non-words by randomly switching trigrams (sequences consisting of three letters) from the words provided. The generated list was read over by two native French speakers independent to the project and independently of each other. They checked to make sure the pseudo-words were pronounceable and with plausible word structure. Real words and any words that resembled real words too closely (e.g., pronounced they sounded like real words) were either removed or further changed by replacing certain letters. Pseudo-words were checked so that none differed by merely a difference of one letter and no pseudo-words shared the first three or the last three letters. This was so as to avoid the appearance of repeats. Further, a list of 100 real words was generated for the targets. These were all 6 letter long, singular nouns of frequency in films between 15-40 as indicated by the Lexique database (New, Pallier, Brysbaert, & Ferrand, 2004).

For Experiment 4b, a list of 226 singular, French nouns, 5-8 letters in length was used. The norms used in Experiment 4a did not yield sufficient number of items. As such to select the items, the Gougehnheim 2.0 norms (New & Pallier, 2001a) were used with the spoken frequency of the words in the range of 10 – 30.

3.2.3 Procedure

3.2.3.1 Experiment 4a

The study was programmed in PsychoPy (Peirce, 2007) and participants completed the entire study on their own on a computer in the presence of the experimenter (see Figure 3.1 for schematic of experimental procedure). Firstly, in the Rating Phase, participants were presented twice with a list of 20 randomly chosen pseudo-words, one item at a time. For one presentation cycle they rated the pseudo-words on pleasantness (on a five-point scale with 1 indicating *not at all* and 5 being *very* pleasant) and on the other presentation cycle they rated them on resemblance to real words (1 being *not at all* and 5 being *very* resembling). The order in which they completed the two judgments was randomized and participants had three seconds to press the key corresponding to their choice for each item. In the subsequent Study Phase participants studied 40 cue-target pairs, each presented for six seconds. The cues consisted of 20 pseudo-words encountered in phase 1 (high familiarity) and 20 were pseudo-words presented for the first time (low familiarity). All targets were real words. This was followed by a Judgment-of-Learning (JOL) Phase in which participants were re-presented with all the cues indicated, by pressing one of two keys, whether *yes* or *no* they would recognize the associated target. Lastly, participants completed a Memory Phase where all the cues were re-presented and participants chose the associated target from among two options (the correct target and another target presented at study), by pressing on the corresponding key.

Participants completed the whole task (all four phases) twice in a blocked design. In one block they were given two seconds to make each JOL (speeded condition) while in the other block they had to wait two seconds from presentation of the cue before making the judgment after which they again had two seconds to respond (delayed condition). Participants were presented with a warning signal 500ms before their time was up, to let them know they should respond. The order in which they completed the conditions was counter-balanced across participants. The trials on which they failed to respond within the given time frame were excluded from analysis. Before each testing session, the cue-target pairs were randomly generated for each participant so that each encountered different items.

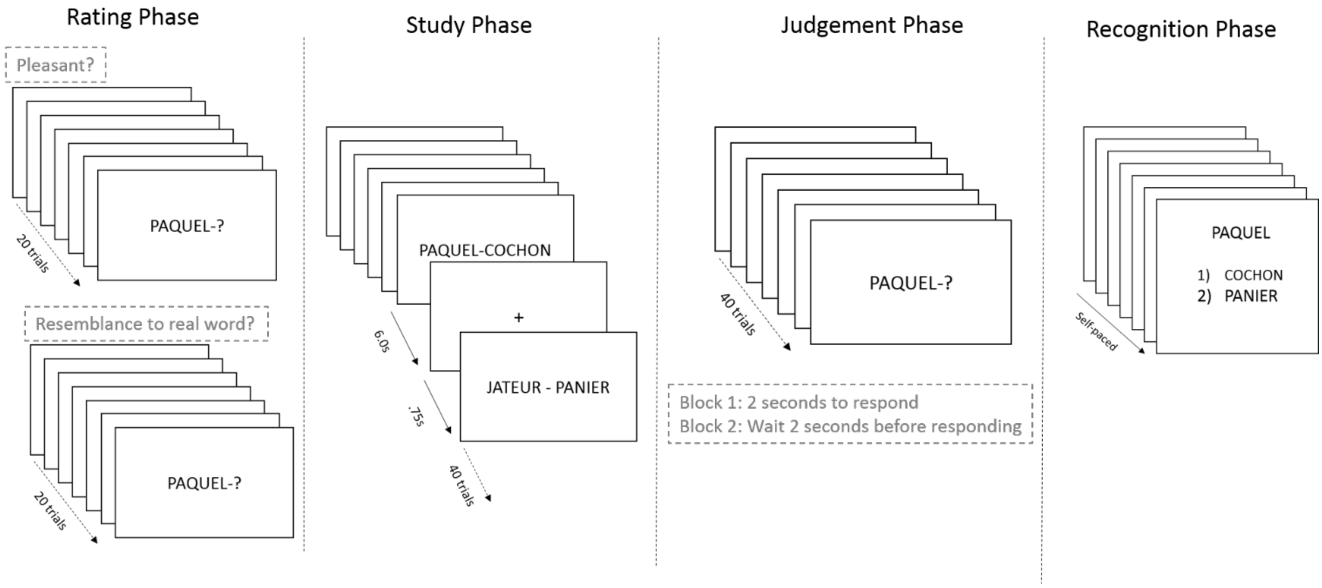

Figure 3.1: Outline of Experiment 4 procedure. *The items presented here are French pseudo-words (cues) paired with real French words (targets; Experiment 4a).*

3.2.3.2 Experiment 4b

The procedure was the same as in Experiment 4a with three exceptions. Firstly, participants studied pairs of real words rather than pairs of pseudo-words and real-words. Secondly, in the Rating Phase, the judgment about resemblance to real-words was replaced by a concreteness judgment (1 representing *very abstract* and 5 representing *very concrete*). We wanted the second rating task in both experiments to encourage deep processing and to relate to the content of the information being processed (rather than its perceptual features). This was so as to ensure the effectiveness of the cue familiarity manipulation. To this end it was not possible to find the same rating task for both real-word and pseudo-word cues and so instead we looked for comparable judgments (abstract/concrete here and resemblance to real-words in Experiment 4a).

3.3 Results

3.3.1 Responding

In the speeded conditions, in Experiment 4a participants took on average 1079ms (*SD* = 215) to respond and in Experiment 4b this was 1083ms (*SD* = 209). Most participants gave a response for all trials. On average, in Experiment 4a participants failed to give a response for 0.3% of trials and this was 1.2% of trials in Experiment 4b.

In the delayed conditions, after the two seconds elapsed and participants were allowed to respond, in Experiment 4a they responded on average 591ms after the deadline (*SD* = 202) and in Experiment 4b they responded in 555ms (*SD* = 177). On average, participants missed 0.6% trials in Experiment 4a and 0.9% trials in Experiment 4b. This shows that overall participants did not struggle with the imposed time deadlines.

3.3.2 Memory Performance

To analyze memory performance, a cue type (pseudo-word, real word) x cue familiarity (high, low) x timing condition (speeded, delayed) mixed ANOVA was carried out on the number of items correctly recognized on the recognition test (see Figure 3.2). There was a main effect of cue type, $F(1, 64) = 4.53$, $p < .05$, $\eta_p^2 = .07$, with better recognition performance in Experiment 4b which employed real word cues than in Experiment 4a which used pseudo-word cues. There was no main effect of timing condition, $F(1, 64) = 1.82$, $p =$

.182, $\eta_p^2 = .03$, or cue familiarity, $F(1, 64) = 1.51$, $p = .224$, $\eta_p^2 = .02$. There was however a cue familiarity x cue type interaction, $F(1, 64) = 7.42$, $p < .01$, $\eta_p^2 = .10$. There were no other significant interactions (all *ps* > .300).

Follow up analyses of the cue familiarity x cue type interaction showed that for high familiarity cues, there was no difference in recognition accuracy between words and pseudowords, $t(64) = 1.03$, $p = .308$, $d = 0.26$, whereas for low familiarity cues, targets paired with real words were better recognized than targets pairs with pseudowords, $t(64) = 3.01$, $p < .01$. All in all, it seems that experimental cue familiarity improved memory performance for targets paired with pseudo-words such that they were remembered as well as targets paired with real-word cues.

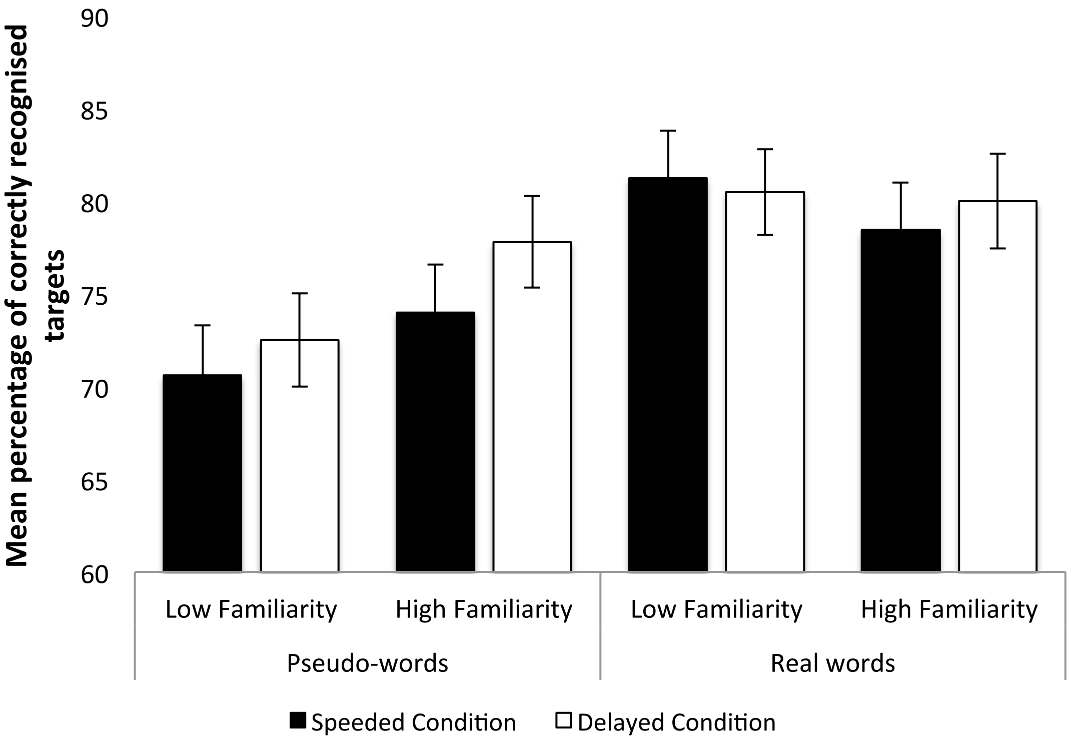

Figure 3.2: Mean percentage of correctly recognised items by condition (fast vs. slow), cue type (pseudo-word vs. real-word) and cue familiarity (high vs. low familiarity). *Error bars indicate standard error of the mean.*

3.3.3 JOL responding

The same ANOVA as above was conducted to analyse the percentage of positive (i.e. *yes*) JOL predictions (Figure 3.3) to explore how the variables of interest impacted responding. Again, there was a main effect of cue type, $F(1,\ 64) = 10.86$, $p < .01$, $\eta_p^2 = .15$, with participants giving overall more positive JOL predictions in Experiment 4b than participants in Experiment 4a. As predicted, the cue familiarity effect was confirmed with participants predicting they will recognize more targets paired with high familiarity than low familiarity cues, $F(1, 64) = 72.40$, $p < .001$, $\eta_p^2 = .53$, across experiments and across conditions. There was also a cue familiarity x cue type interaction showing this difference between positive predictions for high vs. low familiarity cues was greater in Experiment 4a than in Experiment 4b, $F(1,\ 64) = 4.07$, $p < .05$, $\eta_p^2 = .06$. In other words, experimentally induced familiarity had more of an effect on JOL responding when the cues were pre-experimentally novel (pseudo-words) as compared to already known (real words) but both differences were significant ($p < .001$). Further, participants gave more positive JOL responses to low familiarity real-word as compared to low familiarity pseudo-word cues, $t(64) = 4.10$, $p < .001$, $d = 1.03$. There was also a numerical but marginal difference between high familiarity real word and pseudo-word cues, $t(64) = 1.89$, $p = .064$, $d = 0.47$. Contrary to our prediction, there was no effect of timing condition, $F < 1$, nor any further interactions (lowest *p*-value = .110).

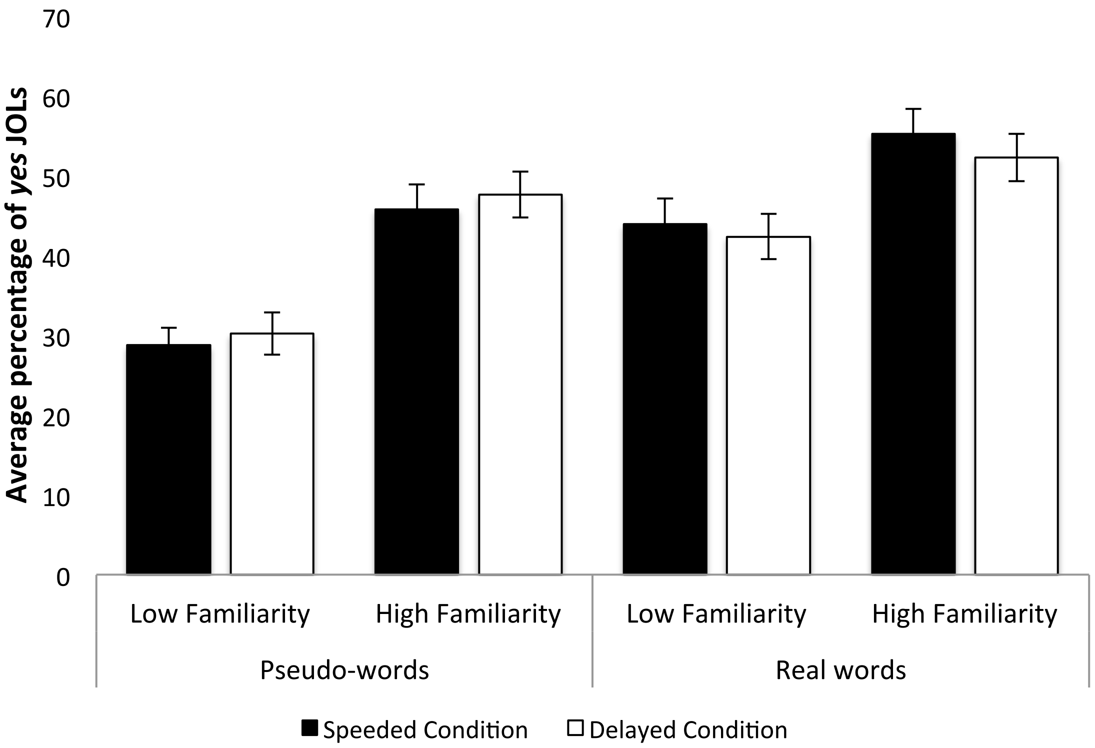

Figure 3.3: Mean percentage of *yes* JOL predictions by condition (fast vs. slow), cue type (pseudo-word vs. real-word) and cue familiarity (high vs. low familiarity). *Error bars indicate standard error of the mean.*

3.3.4 Response bias

To further understand changes in responding, we also analysed criterion or response bias (see Table 3.1), which allowed us to assess responding independent of JOL accuracy. To briefly review concepts introduced in the Introduction; in the context of metacognitive judgments, hits are the number of remembered items given a *yes* JOL prediction and false alarms are the number of not remembered items that were given a *yes* JOL prediction. Bias (see equation 4) is based on the hit rate (ratio of hits to all remembered items) and the false alarm rate (ratio of false alarms to all not remembered items).

The same ANOVA as in previous analyses was used. There was an effect of cue familiarity,$F(1, 64) = 57.87$, $p < .001$, $\eta_p^2 = .90$, a main effect of cue type, $F(1, 64) = 9.04$, p

= .004, $\eta_p^2 = .14$, but no effect of time condition, $F < 1$. Overall, this shows that in both experiments participants responded more liberally when the cue was experimentally more familiar to them.

Table 3.1: Average response criterion by cue type, condition and cue familiarity *(SDs in brackets).*

		Cue familiarity	
Cue Type	Condition	Low	High
Pseudo-words	Delayed	.58 (*.42*)	.16 (*.45*)
	Speeded	.60 (*.41*)	.25 (*.44*)
Real-words	Delayed	.32 (*.41*)	.04 (*.46*)
	Speeded	.25 (*.49*)	-.01 (*.50*)

3.3.5 JOL accuracy

To assess relative monitoring accuracy, we analysed *d'* (see Table 3.2). Using the same ANOVA as described above, we found no effect of cue familiarity, $F(1, 64) = 3.21, p = .078$, $\eta_p^2 = .05$, cue type, $F(1, 64) = 1.74$, $p = .192$, $\eta_p^2 = .03$, timing condition, $F < 1$, or interactions (lowest *p*-value = .201). This means that overall, participants' JOL prediction accuracy was the same across experiments, timing conditions as well as cue types and familiarity.

Table 3.2: Average *d'* by cue type, condition and cue familiarity *(SDs in brackets).*

		Cue familiarity	
Cue Type	Condition	Low	High
Pseudo-words	Delayed	.26 (*.52*)	38 (*.62*)
	Speeded	.14 (*.69*)	.45 (*.70*)
Real-words	Delayed	.41 (*.57*)	.39 (*.49*)
	Speeded	.39 (*.64*)	.51 (*.63*)

3.4 Discussion

The present chapter compared effects of experimental and pre-experimental cue familiarity on delayed JOLs and memory. Pre-experimental familiarity was manipulated by varying the type of cue employed, with Experiment 4a using novel pseudo-word cues and Experiment 4b using real, average frequency words as cues. Experimental familiarity was manipulated through a pre-study exposure to half the cues in a rating task. The results showed that (i) experimental and pre-experimental familiarity aided the creation of novel cue-target associations and these effects interacted with each other, (ii) familiarity increased the rate of positive JOLs given with an interaction between the two types of familiarity, (iii) experimental cue familiarity also influenced response bias with more liberal responding for cues that were experimentally highly familiar as compared to cues that had low experimental familiarity. There were no differences in metacognitive accuracy between items and no effects of time condition.

Firstly, the lack of effect of time condition needs to be addressed. While we did not expect this manipulation to influence memory, we did expect it to impact how cue familiarity and cue type influenced JOL responding. Previous studies have observed that cue familiarity was a strong influence on JOL responding in speeded conditions (Benjamin, 2005, Metcalfe & Finn, 2008b) and exerted diminished (Metcalfe & Finn, 2008b) or no influence (Benjamin,

2005) in delayed conditions. Compared to these other studies, our speeded condition was slower (2000 milliseconds as compared to 750-1000 milliseconds), which likely accounts for this lack of a difference. In other words, it is possible that the effect of cue familiarity on JOLs changes only in that first 1000 millisecond window from cue presentation and then levels off and remains fairly constant. This might be true at least in episodic cue-target paradigms that employ single words, which are fast to process. In contrast, a semantic study, which used general knowledge questions as cues, found changes in the effect of cue familiarity on FOKs between judgments made before two seconds as compared to judgments made after 10 seconds from cue presentation (Koriat & Levy-Sadot, 2001). Similarly, a study that has used triplets of remote associates, asking participants to come up with a fourth word linking to all three presented words observed changes in the impact of cue familiarity on judgments in a much larger time window (Vernon & Usher, 2003). The timing of these influences is very likely task and stimulus dependent; single words are faster to process than sentences or even triplets of words. It is evident that in the context of the current study, the two time conditions were not sufficiently different from each other and the effects investigated in this chapter do not dynamically develop across the time frames employed.

The memory results showed that both types of familiarity influenced the ability to form novel cue-target associations. Memory was better for targets paired with pre-experimentally familiar real words as compared to novel (pseudo-word) cues. Further, experimentally manipulating cue familiarity improved access to the target in the case of pre-experimentally novel cues. More specifically, memory for targets paired with high familiarity cues (presented in the pre-study rating task) was better than for targets paired with low familiarity cues when the employed cues were pseudo-words, with no effect of experimental cue familiarity observed for real-word cues. Further, memory for targets paired with the high familiarity pseudo-word cues was the same as memory for targets paired with real word cues.

The pseudo-word results are consistent with a recent finding that experimentally manipulating familiarity of pre-experimentally novel cues improves the formation of novel cue-target associations (Reder et al., 2016). It is also consistent with Poppenk et al.'s (2010) suggestion that both types of familiarity might provide a similar type of encoding advantage. This is the first study to have directly compared both types of familiarity in an associative task, and to demonstrate the non-additive effects that were observed. Nevertheless, it is necessary to note that the lack of an effect of experimental familiarity on target memory with real-word cues might be due to already high memory performance for these items. In other words, experimental cue familiarity might also aid in associating pre-experimentally familiar cue-target pairs but the advantage of experimentally manipulating familiarity is likely to always be higher (and sometimes exclusive) to pre-experimentally novel items.

Past research has shown that novel information is best encoded when it can be integrated with past experience and knowledge (e.g., Allen et al., 2015; Chase & Simon, 1973; DeWitt et al., 2012; Reder et al., 2013; Staresina & Davachi, 2009), which offers a possible explanation for the current findings. An alternative explanation comes from work on the advantage of deep levels of processing and elaboration to encoding (Craik & Lockhart, 1972), particularly in forming associations (Bower, 1970; Bower & Winzenz, 1970). While we did not instruct participants on how to encode the cue-target pairs, the real-word items offer themselves to techniques such as imagery or the creation of a narrative about the cue-target pairing. This is not possible with the pseudo-word cues. However, it is plausible that with repeated exposure, participants might find ways to disambiguate the pseudo-word cues and maybe even relate them to real words (even though effort was made to ensure this was not possible when creating the items). This would make it possible to use more elaborate encoding strategies when studying a target paired with a highly familiar pseudo-word cue. It would be of interest to compare the present results to a condition where the low familiarity cues were rare or low frequency words rather than pseudo-words. This would allow for the differentiation between the effect of having a cue that carries no meaning and a cue that has low level of familiarity. A possible confound of the present experiments is that this differentiation is not possible.

As predicted, both types of familiarity (experimental and pre-experimental) impacted the percentage of positive (*yes*) JOL predictions given. In other words, participants gave a higher percentage of *yes* JOLs to highly familiar cues as compared to low familiarity cues and more *yes* predictions were also given to real-word as compared to pseudo-word cues. Further, the two effects interacted such that the effect of experimental cue familiarity was bigger for pseudo-word cues. The response criterion analysis further showed that participants were more liberal in their responding for high familiarity cues (both pseudo-word and real-word) as compared to low familiarity cues.

It would be of interest to extend the present paradigm to the study of memory and metamemory in older and clinical populations. Older adults have been shown to be particularly impaired on associative (as compared to item) memory (Old & Naveh-Benjamin, 2008), with this impairment especially evident when asked to make novel associations (Badham & Maylor, 2011; Naveh-Benjamin, Hussain, Guez, & Bar-On, 2003). Similarly, an impairment in forming (novel) associations has been observed in temporal lobe epilepsy (TLE), leading to the suggestion that the memory deficit in TLE could be particularly a deficit in binding (Herfurth, Kasper, Schwarz, Stefan, & Pauli, 2010; Leritz, Grande, & Bauer, 2006; Saling et al., 1993). It would be of interest to investigate the extent to which these types of associative deficits could be ameliorated by the combined contributions of experimental and pre-experimental familiarity. Correspondingly, it would be of interest to investigate how cue familiarity effects would impact metacognitive judgments in a context where the target might be more difficult to retrieve.

In summary, this study showed that both experimental and pre-experimental cue familiarity aid in forming novel associations, confirming and extending previous findings. What is more, memory for targets paired with highly familiar pseudo-words was equivalent to memory for targets paired with real-words, demonstrating that multiple presentations of novel items can facilitate encoding in a manner offered by information stored in long-term memory. Further, both types of familiarity influenced metacognitive judgments, with an increase in *yes* JOLs with both experimental and pre-experimental familiarity manipulations but these effects interacted such that there was no effect of experimental familiarity on pre-experimentally familiar cues. In contrast to the memory findings, the effects of cue familiarity and cue type on JOL responding were additive. It is clear that while manipulations that affect memory are also likely to influence metamemory, the effects do not have to map directly onto each other. Further, cue effects in delayed JOL have been seen as primarily prominent in initial, pre-retrieval processes as exhibited in highly speeded judgments (within 1000 milliseconds of cue presentation; e.g., Benjamin, 2005; Metcalfe & Finn, 2008b). Here we extend the role of the cue to judgments made later and show that different types of cue familiarity (experimental and pre-experimental) can influence delayed JOLs.

CHAPTER 4

INSIGHTS FROM JUDGMENT-OF-LEARNING JUSTIFICATIONS

4.1 Introduction

Metacognitive judgments are understood as corresponding to quantity and quality of some (internal) evidence gathered toward the judgment being made (e.g., ease of reading as evidence that an item has been sufficiently learned and will be later remembered; Rhodes & Castel, 2008) and reflecting the probability that the given judgment is correct (Kepecs & Mainen, 2012). Correspondingly, Chapters 2 and 3 showed that varying levels of access to target related information and cue familiarity influences delayed JOLs. As noted in the discussion of Chapter 2, one of the worries in metacognitive research is that when we ask participants about access to particular type of information (such as memory for target spatial location) or when we manipulate the strength of certain features (such as cue familiarity), we are making these features particularly salient to participants. The outstanding question then is whether we would observe the same effects if we did not explicitly manipulate or probe these factors and rather only relied on their natural variation across items in any given task. This chapter investigates what metacognitive judgments represent by evaluating how participants spontaneously construct and justify their delayed JOL confidence. Participants provided written reports alongside their JOLs and we used natural language processing techniques to characterize the type of information and explanation that differentiate one JOL confidence from another, and to quantify the extent to which any two JOLs are justified with reference to different types of evidence. We also manipulated whether participants gave JOLs on a confidence scale (Experiment 5) or as a binary judgment (Experiment 6). This allowed us compare the two response formats directly and evaluate how they relate to each other in terms of how they are constructed.

This study draws on research investigating retrospective confidence in contents of memory retrieval, which has established that probing participants for explanations and justifications of their answers is a powerful tool for characterizing processes underlying cognition and metacognition. For example, Koriat et al. (1980) asked participants to list reasons for and against their chosen answer to a general knowledge question. They observed that confidence was influenced by the amount of evidence accessed in support of the given answer, lending support to the idea that confidence is a result of a process of evaluation of different sources of evidence. More recently, Selmeczy and Dobbins (2014) asked participants to justify why they were confident (or not) in their judgments on a recognition task (calling an item *old* or *new* i.e. previously studied or seen for the first time). Their analyses of the content of these responses showed a pattern of results supporting dual-process accounts of recognition memory (see Yonelinas, 2002); for example, the presence of ‘remembering’ characterized high confidence *old* responses in contrast to medium confidence responses which were more likely to contain references to ‘familiarity’. In other words, this quantitative analysis of subjective reports lent support to one side of an on-going debate in recognition memory. Furthermore, these results were obtained without explicit instructions or theory-laden manipulations from the experimenters, who did not highlight specific experiences or types of evidence for participants to focus on. Overall, these studies (see also Gardiner, Ramponi, & Richardson-Klavehn, 1998; Urquhart & O’Connor, 2014; Williams, Conway, & Moulin, 2013) give credibility to the idea that much can be learned from asking participants to explain their metacognitive judgments and experiences even though it remains uncommon practice. In the present study we adopted and developed the analytical approach pioneered by Selmeczy and Dobbins (2014) to gain insight into processes underlying delayed JOLs.

Based on results of the previous chapters and the literature discussed, we expected participants to reference both cue familiarity and target accessibility in their justifications. More importantly, we were interested in observing how the type of evidence referenced mapped onto the confidence expressed. Further, we wanted to observe whether the pattern of justifications for confidence responses would map onto justifications of binary responses.

There is an underlying assumption in JOL confidence analysis that the confidence responses can be split into binary (*yes* and *no*) predictions. This is implicit in the use of calibration measures, which assume that confidence responses correspond to a prediction of likelihood of future retrieval (e.g., Finn & Metcalfe, 2007; Koriat, Sheffer, & Ma'ayan, 2002; Serra & England, 2012). This corresponds to the idea that low confidence JOL predictions should probabilistically equate to a rejection of future retrieval (i.e. 40% predicted success rate means greater likelihood of failure). This has meant that some have explicitly suggested to split the confidence scale down the middle to enable calculations of statistics such as *d'* (see for example Masson & Rotello, 2009). And yet recent results suggest that JOL confidence responses and binary responses might not be equivalent and can lead to different pattern of results (Hanczakowski et al., 2013).

The comparison of confidence to binary responses does not have only methodological consequences but also holds theoretical interest. To review the theoretical developments in the delayed JOL literature, early theories have focused on explaining JOLs as a result of single process (target retrieval) evaluations (e.g., Nelson & Dunlosky, 1991). In this view it was assumed that participants accrue one type of evidence (the degree to which the target is accessible) toward their JOL—the more evidence they collect, the higher their JOL. According to this view, different JOLs (e.g., 60% as compared to 80%) merely expressed different degrees of access to the target. Recently, an alternative two-stage view proposed a quick pre-retrieval stage driven by cue-familiarity followed by an effortful memory search (target accessibility evaluation) stage (Benjamin, 2005). Metcalfe & Finn (2008b) further elaborated this view, suggesting the first stage can result in (i) a quick “don’t know” decision driven by lack of cue familiarity (expressed as responding with the lowest point on the JOL scale) or (ii) the initiation of the second effortful retrieval stage. In this case, the prediction is that there are qualitatively different processes that underlie the lowest confidence JOL (i.e. 0%) and distinguish it from all others. More specifically, it is a cue-driven evaluation as compared to a target-based judgment. If this is true, we would expect participants to refer to these types of evidence in their justifications and to observe a qualitative difference in the evidence favoured at different levels of the JOL scale.

An alternative two-stage view has focused on explaining JOLs as consisting first of a *yes* or *no* judgment, directly followed by an assignment of confidence. Dunlosky et al. (2005) observed that when participants were asked to make a confidence judgment about the accuracy of their JOL prediction (a second-order judgment, SOJ), a plot of the SOJ magnitude against JOL confidence resulted in a U-shaped function. In other words, participants were most confident in the predictions that lay on the extremes of the JOL scale and as their JOL confidence moved toward the mid-ranges, participants became less confident in the accuracy of their retrieval predictions. Dunlosky et al. (2005) interpreted the anchor of the SOJ function (i.e. the function’s minimum, where participants were least confident in the accuracy of their JOLs) as the point where *yes* and *no* predictions diverge. This would imply that the low end of the JOL scale is interpretable as *no* predictions, and the high end as *yes* predictions. Dunlosky et al. (2005) speculated that *yes* and *no* predictions are driven by different evidence evaluation processes, although their exact nature has not been specified. Nonetheless, this framework suggests a point of divergence within the range of JOLs available to participants with the point placed in the mid-regions of the scale. More specifically, it suggests that *yes* and *no* predictions are qualitatively different and that a point can be located on the confidence scale where they diverge.

While the models described above are not irreconcilable, they do lead to a somewhat different pattern of predictions. Primarily, they both suggest there is an underlying point of divergence in the JOL scale either side of which the scale is characterized by different processes. Metcalfe and Finn (2008b) place that point on the lowest ends of the scale and describe it in terms of the information evaluation processes that change at that point while Dunlosky et al. (2005) place it in low to mid ranges (see also Serra & England, 2012) and describe it in terms of a *yes/no* distinction. This *yes/no* distinction along the center of the scale has similarly been hypothesized elsewhere (Hanczakowski et al., 2013) and is consistent with confidence interpretations described above. Notably, the speculated *yes/no* distinction in confidence judgments has not yet been demonstrated and its potential relationship to cue as compared to target related processes is poorly understood.

The aim of this chapter was to evaluate the types of evidence that characterize confidence and binary JOLs and to compare these against each other, thus evaluating methodological and theoretical assumptions made in the literature. Across two experiments, participants completed a standard JOL task with cue-target word pairs. In Experiment 5 participants made JOL predictions on a 6-point numeric confidence scale (0-20-40-60-80-100%) whereas in Experiment 6 participants made first a binary *yes*/*no* JOL prediction followed by a three-point verbal confidence judgment made about that prediction (*sure-maybe-guess*). In both experiments there was a total of six JOL response options and participants provided written justifications on a subset of their JOLs. They were not given any instructions on how to write their justifications nor did we manipulate any variables known to influence JOL confidence. Thus, we assessed how participants arrive at JOL confidence independently and spontaneously without making any one source of information (e.g., cue familiarity) more salient than others. The general procedure and majority of methods adopted to analyze the text data were modeled on Selmeczy & Dobbins (2014). Across two experiments we examined: (i) how participants justify their JOLs; (ii) to what extent are such justifications characterized by cue and target references; (iii) whether there is an underlying *yes*/*no* distinction in numeric JOL confidence responses.

4.2 Method

4.2.1 Participants

All participants that took part were affiliated with the University of Leeds (students and staff) with 54 participants (13 men; mean age = 23.4; *SD* = 7.4) in Experiment 5 and 73 participants (12 men; mean age = 27.5, *SD* = 10.7) in Experiment 6. In Experiment 5, two participants were excluded, both for not following instructions (one for using only 0% and 100% judgements, the other because her written responses referred to multiple cue-target pairs instead of the pair preceding the written report). This left 52 participants in the analysis for Experiment 5 (13 men, mean age = 22.5, *SD* = 6.2). In both experiments, participants either received course credit or £5 as reimbursement. The study was granted ethical approval by the School of Psychology Ethics Committee, University of Leeds, UK.

4.2.2 Stimuli

For each participant, the studied items were randomly selected from a list of 628 common, singular English nouns (5-6 letters long) taken from the English Lexicon Project (minimum log Hyperspace Analogue to Language frequency 8.02; Balota et al., 2007). Each participant was exposed to a unique set of 90 cue-target pairs (45 in each of the two experimental blocks).

4.2.3 Procedure

The study was programmed using PsychoPy (Peirce, 2007) with all participants completing the task individually on a computer, in the presence of the experimenter. In both experiments, participants completed two identical blocks consisting of three consecutive phases (see Figure 4.1); the task was repeated so as to collect sufficient number of JOL justifications. In each block participants: (i) studied 45 cue-target pairs-presented for six seconds each with a fixation cross between all trials; (ii) were presented with the cue of each pair, and gave a JOL predicting recognition performance for the target on the subsequent memory test; and (iii) completed a forced choice recognition test where, on presentation of each cue, they selected the cue-matched target from two words (both options were targets from the study). All choices were made by pressing a key corresponding to the confidence response or target.

The only difference between the two experiments was in the JOL stage (part ii of the procedure). In Experiment 5, participants gave their JOLs on a six-point numeric confidence scale (0-20-40-60-80-100%). In Experiment 6, participants first gave a binary y*es/no* response indicating whether they would recognise the target, followed by a three-point verbal confidence judgment (*sure-maybe-guess*) relating to the *yes*/*no* response. In both experiments there were 6 distinct response options participants could give.

On a subset of the judgment trials, immediately after giving a JOL, participants justified the previously rendered JOL using a written, keyboard-entered response. Over the two blocks participants could give a maximum of 18 justifications (nine per block)—three per response option. More specifically, no questions were asked on the first five trials of either block. After that, requests for written justifications were spread out throughout the judgment task as follows. If the maximum number of justifications was reached for a given JOL response type, no more justifications were asked for that response option. Participants would not be asked for any written responses for the two trials following a justification, though this enforced gap reduced over the course of the block (there was no enforced justification gap for the last 10 trials). Some participants therefore gave fewer judgments than others, especially since some participants would use some JOL responses less than others. On average participants gave 15.4 justifications in Experiment 5 (*SD* = 1.9) and 12.7 justifications in Experiment 6 (*SD* = 2.7).

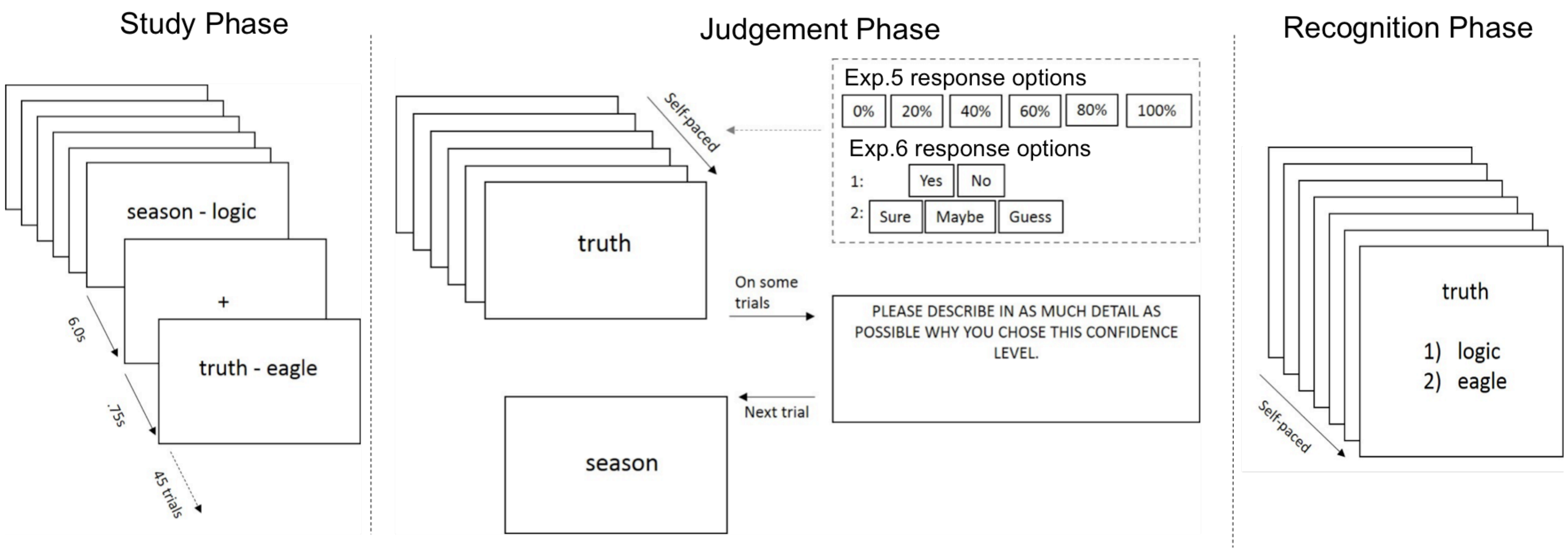

Figure 4.1*:* Schematic of Experiment 5 and 6 procedure. *The three phases together constitute one experimental block. Participants completed two blocks, with a new set of items in each. In the Judgment Phase, participants gave a JOL with variation in response format across experiments. In Experiment 5, participants indicated their numeric confidence in one response whereas in Experiment 6, they gave a binary judgments (yes/no) before indicating their verbal confidence in this judgement. On a subset of trials participants were asked to also explain why they gave the particular JOL prediction on the preceding trial.*

4.2.4 Text analysis methods

4.2.4.1 Text data pre-processing

Before any text analysis was carried out, we corrected spelling mistakes in the text and removed articles (*a* and *the*). We also removed justifications where participants used it to explicitly indicate they wanted to change the JOL response they had given (in total, three justifications in Experiment 5, six justifications in Experiment 6). In all of the reported analyses, we aggregated the descriptive reports for each JOL confidence level and response type across participants for comparison. At least 100 justifications were collected per JOL type (see Table 4.1 for number of justifications collected per JOL response category).

4.2.4.2 Latent Semantic Analysis (LSA)

LSA is a technique by which one can evaluate the semantic relationship between a single term and a text document. Drawing on singular value decomposition (closely related to factor analysis), LSA creates a mathematical (matrix/vector) representation of a large body of text, mapping the semantic relationships between single words and sets of words. This mapping relies on frequency of co-occurrence but also on a weighting function that takes into account the 'importance' of a term to a given text (see Landauer, Foltz & Laham, 1998 for more detail). LSA that has been trained on a relevant corpus of texts (e.g., general or subject specific) to create this representation, also called semantic space, can then be applied to new examples to compute their semantic relationship. The subsequent classification of semantic similarities between new examples very closely imitates humans (e.g., Laham, 1997). The online LSA tool (available at http://lsa.colorado.edu/) offers a semantic space that has been trained on 'general reading' corpus with 300 factors (Dennis, 2006). We used this to classify the semantic similarity between each justification and the cue-target pair it was written in response to. More specifically, we computed an LSA score between the cue and the justification and compared it against the LSA score computed between the target and the justification. The toolkit returns a cosine value for each comparison; as such the range of output values is -1 to 1, with 0 or lower interpreted as no semantic relationship. Following Wandmacher, Ovchinnikova, and Alexandrov (2008), we set negative LSA values to 0 since in this context we could not interpret a justification and a studied item (cue or target) as being more dissimilar than 'not similar at all'. If, for example, a justification for a given JOL response type is more likely to refer to the cue than the target (e.g., "I cannot remember studying the word truth" where 'truth' is the cue) then the LSA value should be higher for the cue-justification as compared to the target-justification comparison. This enabled us to assess whether any JOL category was characterized by referring more to the cue or the target, as predicted by Metcalfe and Finn's (2008b) two stage JOL account.

4.2.4.3 Word frequency analysis (n-grams)

An n-gram is a continuous series of words found to occur within a text (n = 1, 2, 3 are referred to as uni-grams, bi-grams and tri-grams respectively). To compare sets of texts (in this case, justifications) the frequency of occurrence of each n-gram is counted across all justification texts. To account for some participants writing more than others (and possibly repeating themselves), we restricted the analysis so that each JOL justification could contribute a maximum of 1 to any given n-gram count. For any given n-gram (e.g., "do not remember") we could thus compute the total number of justifications that contained it for each JOL category.

In previous experiments analysing *n*-grams (Selmeczy & Dobbins, 2014; Urquhart & O'Connor, 2014), only two categories were ever compared against each other. This was done using a binomial test, computing a *p*-value for the proportion of occurrence of the given *n*-gram under one response category assuming a binomial distribution with the *p*-parameter of 0.5. This allowed for the examination of whether the *n*-gram was significantly more likely to appear in justifications for one response category or whether the probability of it occurring in texts justifying either response category was equal. Since in this study we had six response categories in each experiment, we adapted the binomial test to account for probability of success of 1/6. In other words, for each JOL category, we computed whether that proportion of occurrence (out of all occurrences) was significantly higher than that *n*-gram having equal probability of occurrence in all JOL categories.

This analysis allowed the isolation of simple phrases that were most likely to be used in justifying one JOL response type as compared to all others. Where LSA focused on semantic similarity between the studied items (cue and target) and the justification texts, *n*-gram analysis examined whether different phrases (e.g., relating to familiarity as compared to retrieval success) would differentiate different JOL response categories. Rather than analysing information specific to each trial (i.e. whether participants named or referred to the studied items), this analysis enabled the extraction of general phrases that held true across trials, irrespective of what the studied cue or target were. In this way the *n*-gram analysis complemented, and helped to further explicate, the LSA results.

4.2.4.4 Classification analysis (Support Vector Machine [SVM])

SVM is a machine-learning algorithm commonly used in text classification. Here we employed it as a tool for quantifying the extent to which different JOL responses differed from each other. If there are highly distinct features that separate one category from another (such as reference to different types of processes), then the SVM would pick up on this and classification of future examples would be highly accurate. On the other hand, if the differences were merely of degree (e.g., different levels of target access), then the classification of future examples would be low.

To carry out SVM analysis, we represented each written justification as a vector where each vector component corresponded to a *uni*-gram, *bi*-gram or *tri*-gram, with 0 denoting its absence in the given justification text and 1 denoting its presence. We included all *n*-grams as this allowed us to account for individual word usage as well as word combinations, which carry specific semantic meaning. For example, the *uni*-grams 'not', 'remember' and 'confident' could only be coded as present once which would mark the texts 'I am confident I will not remember' and 'I am not confident but might remember' as the same while including the bigrams 'not remember' and 'not confident' avoided this problem. Each *n*-gram thus constituted an input feature and each text was represented as a vector of features while the output was the JOL category the given vector belonged to (e.g., 0%). In principle, an SVM looks for a 'decision boundary' or a line that separates the two sets of data being compared so that the distance between the boundary and any point of any class is the biggest it can possibly be—that is why it is called a maximum-margin classifier (Hamel, 2009). Once an SVM has been trained it can be used to classify new data which will be assigned either of the categories the SVM has been trained on, based on which side of the margin it falls on.

The SVM analysis was implemented with scikit-learn, an open source toolkit developed for Python (Pedregosa et al., 2011). To compare two JOL response categories (e.g., 0% vs. 20% JOL), the justification responses for both were labeled and combined. We trained the classifier on a randomly selected half of the combined data with a linear kernel and a cost value of 0.10 and tested it on the other half. Once the classifier was trained, it was then used to classify the remaining half of the data, and its performance was evaluated by its ability to distinguish correctly what JOL a given text was written for. The JOL confidence models described in the Introduction both speculate a divergence on the confidence scale with regards to the processes that drive the judgment. A difference in processes relied upon (i.e. a qualitative difference) should lead to high classification accuracy whereas differences merely of degree (i.e. quantitative differences) should lead to low classification accuracy due to low likelihood of distinct, differentiating features.

4.3 Results

4.3.1 Memory and JOL responses

In Experiment 5, participants correctly recognised 84.7% (*SD* = 11.6) targets on the final memory test. In Experiment 6 they correctly recognised 86.2% (*SD* = 12.2) of targets. Memory performance did not differ between the two experiments, $t < 1$.

Since we did not manipulate anything, we did not expect accuracy to differ. Nevertheless, we report *d'* and AUC for interest. In Experiment 1, average AUC for block 1 (*M* = .674, *SD* = .154) was the same as mean AUC in block 2 (*M* = .713, SD = .144), *t*(51) = 1.91, *p* = .062, *d* = .03. In Experiment 2, *d'* in the first block (*M* = .717, *SD* = .487) was the same as *d'* in the second block (*M* = .705, *SD* = .605), $t < 1$. See Figure 4.2 for the mean proportion of trials each JOL category was used and Table 4.1 for the number of written justifications collected per JOL category.

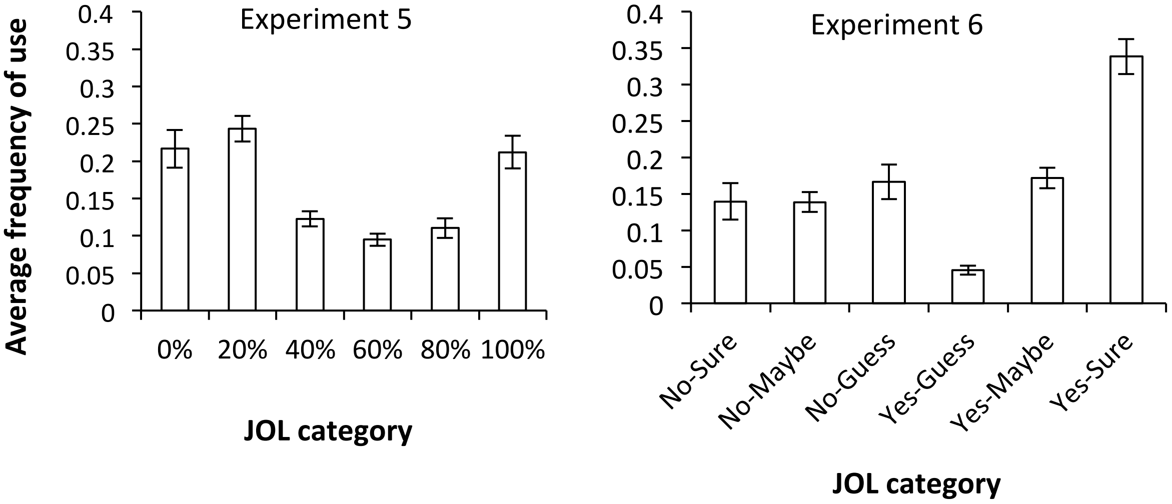

Figure 4.2*:* Mean percentage of trials in each JOL category by experiment. *Error bars indicate standard error of the mean.*

Table 4.1: Number of justifications collected in each JOL category by experiment.

Exp5	0%	20%	40%	60%	80%	100%
	127	146	134	120	132	137
Exp6	No - Sure	No - Maybe	No - Guess	Yes - Guess	Yes - Maybe	Yes – Sure
	102	177	137	102	195	205

4.3.2 Latent Semantic Analysis (LSA)

Metcalfe & Finn (2008b) proposed that the lowest point on the JOL confidence scale should reflect the result of a cue-evaluation stage whereas all other JOL levels should correspond to target access evaluations. We used LSA to evaluate whether for each JOL response type, participants were more likely to refer semantically to the cue or the target in their justifications (or neither). For each trial with a JOL justification, we computed an LSA value between the cue and the written justification and compared it against the LSA value computed between the target and the justification. Because the written justifications refer to specific memories, one could expect that overall the semantic similarity scores would be fairly low. However, if participants refer specifically to the cue or the target term (or information relating to them) this would increase the score. Additionally, because LSA has been shown to successfully map meaning (Laham, 1997), this would be true even when participants did not directly refer to the cue or the target but, for example, reported partial semantic information about them. We used paired-samples *t*-tests to compare the cue-justification and target-justification LSA scores for each JOL response category (e.g., 0% JOL confidence) to analyse whether the JOL justifications were more likely to refer to the cue or the target term. The LSA scores range from 0 (no relationship) to 1 (high semantic relationship). This analysis was done for both Experiment 5 and 6 separately with the results reported in Table 4.2.

Table 4.2: Mean cue-justification and target-justification LSA scores in each JOL category by experiment.

	JOL	Cue	Target				
Exp.	category	LSA score	LSA score	*t*-value	df	*p*-value	*d*
5	0%	.21 (.14)	.17 (.12)	2.29	120	.024*	0.30
	20%	.20 (.13)	.17 (.10)	2.42	143	.017*	0.27
	40%	.20 (.11)	.19 (.11)	0.55	133	.581	0.07
	60%	.17 (.12)	.19 (.13)	1.58	119	.118	0.19
	80%	.20 (.13)	.21 (.14)	0.39	129	.700	0.05
	100%	.21 (.12)	.24 (.14)	2.09	134	.039*	0.22
6	No - Sure	.08 (.12)	.06 (.11)	1.14	98	.259	0.12
	No - Maybe	.08 (.12)	.08 (.13)	0.38	171	.704	0.03
	No - Guess	.11 (.14)	.08 (.13)	2.79	133	.006*	0.25
	Yes - Guess	.14 (.17)	.09 (.13)	2.64	101	.010*	0.31
	Yes - Maybe	.10 (.13)	.09 (.13)	1.03	192	.302	0.08
	Yes - Sure	.14 (.17)	.16 (.19)	1.37	202	.172	0.11

Note. (standard deviations appear in parentheses). Results of paired-samples t-tests comparing the cue and target LSA scores within each JOL category are reported. *s indicate significance at an alpha threshold of .05.

The results of the LSA revealed that in Experiment 5, the 0% and 20% JOL confidence level justifications were more likely to semantically refer to the cue than the target. On the other hand, the 100% level was more likely to refer to the target than the cue. The pattern of results of Experiment 6 showed it was the *guess* responses (for both *no* and *yes* predictions) that were more likely to refer semantically to the cue rather than the target term. These results demonstrate that participants rely on both cue and target related information in justifying their JOLs and that these two types of processes provide a useful framework for differentiating different types of JOL predictions. To understand more precisely whether the cue-references were the same or differed between the different JOL responses we turned to word-frequency analysis.

4.3.3 Word-frequency analysis

The next step in the analyses was the examination of unique phrases that differentiated one JOL response from all others. This allowed us to determine whether the cue references in JOL justifications were of the same character (e.g., expressing lack of cue familiarity) or whether they relied on the cue term differently (e.g., cue familiarity characterizing 20% whereas its absence characterizing 0% JOL). Further, whereas LSA only tracked semantic similarity, participants could express lack of cue familiarity without actually naming the cue itself (e.g., "This cue is not familiar"). Compared to LSA, *n*-gram analysis thus allowed us to capture these types of phrases and extract meaningful patterns of expression across trials that were significantly more likely to occur for one type of JOL response as compared to others. For example, we expected to see an increase in recollection-specific terminology with increases in JOL confidence as well as greater use of intensity modifiers indicating greater certainty of access.

To constrain the number of *n*-grams analysed, we focused only on *bi*-grams and *tri*-grams with a minimum total occurrence of 10 (stricter than previous analyses which have included *uni*-grams and used lower median occurrences). We only reported *tri*-grams and *bi*-grams reaching significance at $p < .05$ (Table 4.3 reports *n*-gram analysis results for Experiment 5, Table 4.4 for Experiment 6). For each JOL, the analysis extracted phrases that occurred significantly more often than would be expected if the phrase was used equally across all JOL responses. Notably, this does not preclude the possibility that certain phrases might have significantly higher proportion of occurrence than 1/6 for two JOL category responses (e.g., if they never occurred for any other response) and thus allows for extraction of similarities (e.g., are there certain phrases that characterize *no* predictions that are never employed in *yes* predictions) as well as the expected characterization of differences.

Table 4.3: ***n*-gram analysis results for Experiment 5.**

JOL	*n*-gram	Count	Total	Proportion	*p*
0%	not remember this	8	11	.73	<.001
	remember seeing this	13	30	.43	<.001
	remember what word	6	11	.55	<.001
	do not remember	39	66	.59	<.001
	seeing this word	13	31	.42	<.001
	I do not	43	79	.54	<.001
	remember this word	10	26	.38	.007
	I cannot remember	17	45	.38	<.001
	not remember seeing	25	32	.78	<.001
	cannot remember what	9	19	.47	<.001
	cannot remember word	5	11	.45	.025
	do not	58	114	.51	<.001
	not remember	42	73	.58	<.001
	that word	6	16	.38	.038
	have no	7	11	.64	<.001
	this word	31	99	.31	<.001
	word at	6	10	0.6	.002
	I do	43	81	.53	<.001
	seeing this	14	33	.42	<.001
	at all	18	23	.78	<.001
	remember seeing	35	105	.33	<.001
	cannot remember	34	88	.39	<.001
	I cannot	22	78	.28	.009
20%	seeing word but	7	10	.70	<.001
	be able to	12	37	.32	.024
	do not think	8	14	.57	<.001
	not think I	7	13	.54	<.001
	vaguely remember seeing	8	10	.80	<.001

	but I cannot	6	14	.43	.019
	but cannot remember	7	17	.41	.015
	I am not	11	30	.37	.011
	do not really	7	10	.70	<.001
	what it was	7	17	.41	.015
	remember seeing word	14	42	.33	.011
	do not remember	18	66	.27	.030
	I do not	26	79	.33	<.001
	I cannot remember	14	45	.31	.015
	I remember seeing	13	39	.44	.009
	what it	9	23	.39	.009
	not confident	6	12	0.5	.008
	be able	12	37	.32	.024
	not really	8	11	.73	<.001
	not think	9	15	.60	<.001
	am not	11	33	.33	.017
	I cannot	24	78	.31	<.001
	word so	5	11	.45	.024
	really remember	10	15	.67	<.001
	seeing word	17	48	.35	.001
	vaguely remember	11	14	.79	<.001
	able to	12	39	.31	.029
	might be	6	10	.60	.002
	with it	6	15	.40	.027
	I do	26	81	.32	<.001
	do not	39	114	.34	<.001
	not remember	20	73	.27	.018
	cannot remember	24	88	.27	.014
	remember seeing	34	105	.32	<.001
40%	think I remember	6	16	.38	.038

	word but cannot	7	11	.64	<.001
	word but I	7	13	.54	<.001
	if I saw	7	14	.50	<.001
	I remember seeing	17	39	.44	<.001
	think I could	6	13	.46	.013
	remember word but	5	11	.45	.025
	I think I	15	52	.29	.025
	word and	9	29	.31	.046
	word it	5	12	.42	.036
	word I	7	18	.39	.021
	second word	8	22	.36	.021
	to recognise	7	18	.39	.021
	I could	20	63	.32	.003
	but I	20	71	.28	.016
	if I	13	29	.45	<.001
	word but	25	58	.43	<.001
	I may	8	19	.42	.008
	but cannot	12	29	.41	.001
	cannot recall	8	19	.42	.008
	I think	27	85	.32	<.001
	recognise it	10	28	.36	.018
	think I	26	82	.32	<.001
	I remember	41	176	.23	.026
	remember seeing	26	105	.25	.035
60%	but I am	7	18	.39	.021
	I think I	21	52	.40	<.001
	remember making	7	13	.54	<.001
	could recognise	5	10	.50	.015
	I might	6	16	.38	.038
	I feel	13	35	.37	<.001

	and I	9	20	.45	<.001
	I am	19	71	.27	.037
	think I	27	82	.33	<.001
	feel I	6	13	.46	.013
	I can	13	45	.29	.042
	pair word	8	15	.53	<.001
	it but	6	15	.40	.027
80%	I remember word	10	28	.36	.018
	I am pretty	9	14	.64	<.001
	in my head	7	16	.44	.010
	one of	5	11	.45	.025
	am pretty	9	14	.64	<.001
	pretty sure	5	10	.50	.015
	it was	19	70	.27	.024
	I remember	43	176	.24	.008
	in my	10	31	.32	.028
	I associated	6	14	.43	.019
	my head	7	19	.37	.028
100%	I can remember	11	26	.42	<.001
	link between	8	18	.44	.005
	as I	7	19	.37	.028
	thought of	5	10	.50	.015
	it is	9	25	.36	.026
	can remember	12	31	.39	<.001
	I can	14	45	.31	.015
	I made	10	27	.37	.009

Note. A count of occurrences of each *n*-gram in justifications for the corresponding JOL category are reported along with total number of occurrences, proportion of occurrence and p-value computed using the binomial test.

The *n*-gram analysis results presented in Table 4.3 show the 0% JOL confidence level was characterized by an inability to remember ("do not remember") and could be interpreted as expressing lack of cue familiarity as participants indicated they cannot even remember having seen the presented word at study ("not remember seeing"). The 20% JOL confidence level on the other hand was characterized by a vague sense of cue familiarity ("vaguely remember seeing [word]") accompanied by a lack of recollection for the target term ("but cannot remember"… "what it was"). While the LSA results revealed that the 0% and 20% JOL confidence levels were more likely to refer to the cue than the target term semantically, the *n*-gram analysis showed they nevertheless differed from each other in whether the cue term was said to be remembered. The 40% JOL also referenced cue familiarity suggesting the role of the cue in JOLs is not isolated to lowest confidence responses when it is not familiar but can in itself provide a degree of evidence when the target cannot be accessed. Indeed, justifications for the 40% and 60% JOL confidence levels expressed feelings of possible target access ("I think I could recognise but cannot recall") whereas the 80% JOL confidence level started bringing in language of certainty ("pretty sure") and memory for associations ("I associated"). Unsurprisingly, the 100% JOL expressed memory for the target term ("I can remember"). All in all, this pattern of descriptions fits with Metcalfe and Finn's (2008b) suggestions that a lack of cue familiarity leads to a 0% JOL confidence response whereas, when the cue is recognized, the JOL confidence increases with increase in target access. The results further demonstrated that the role of the cue does not stop after that initial stage and is carried as evidence through to the target access stage.

Table 4.4: ***n*-gram analysis results for Experiment 6.**

JOL	*n*-gram	Count	Total	Proportion	*p*
No-Sure	do not remember	28	78	.36	<.001
	I do not	37	93	.40	<.001
	remember this word	11	31	.35	.012
	cannot remember seeing	7	12	.58	<.001
	not remember this	9	12	.75	<.001
	word at all	10	21	.48	<.001
	not remember seeing	12	32	.38	<.001
	do not even	11	12	.92	<.001
	not even remember	11	11	1.00	<.001
	do not	53	137	.39	<.001
	this word	32	120	.27	.007
	not remember	33	89	.37	<.001
	even remember	11	11	1	<.001
	not recognise	5	11	.45	.025
	I do	37	99	.37	<.001
	have no	5	12	.42	.036
	no idea	6	11	.55	.004
	word at	10	25	.4	.005
	not even	11	13	.85	<.001
	at all	16	34	.47	<.001
	remember seeing	24	90	.27	.016
No-Maybe	able to recognise	6	13	.46	.013
	might be able	8	15	.53	<.001
	not sure if	5	11	.45	.025
	be able to	23	66	.35	<.001
	I cannot remember	23	78	.29	.005
	that I would	5	10	.50	.015
	am not sure	7	20	.35	.037

if I saw	6	13	.46	.013
I might be	7	15	.47	.007
I would recognise	7	20	.35	.037
may be able	7	11	.64	<.001
do not remember	20	78	.32	<.001
I would remember	7	10	.70	<.001
cannot remember what	8	19	.42	.008
cannot remember	41	144	.28	<.001
word but I	9	28	.32	.039
I saw	9	27	.33	.034
associated with	8	24	.33	.048
of two	6	14	.43	.019
now but	5	10	.50	.015
not sure	18	61	.30	.014
be able	23	66	.35	<.001
would remember	7	10	.70	<.001
to recognise	6	15	.40	.027
recognise it	13	37	.35	.006
to pair	5	11	.45	.025
what word	9	23	.39	.009
would be	11	34	.32	.021
for this	6	14	.43	.019
I feel	10	24	.42	.003
sure if	5	11	.45	.025
to me	8	12	.67	<.001
I cannot	26	103	.25	.024
if I	29	49	.59	<.001
it if	7	13	.54	.002
to mind	6	14	.43	.019
I may	13	21	.62	<.001

	would recognise	8	21	.38	.016
	remember it	14	49	.29	.034
	I would	27	74	.36	<.001
	I might	14	41	.34	.006
	may be	8	18	.44	.005
	able to	23	67	.34	<.001
	it but	10	22	.45	.002
	remember what	12	34	.35	.009
	might be	10	27	.37	.009
	not remember	22	89	.25	.047
	but I	26	92	.28	.005
	word but	19	72	.26	.038
No-Guess	to guess	6	16	.38	.038
	word so	9	24	.38	.012
	do not remember	25	78	.32	<.001
	I do not	27	93	.29	.046
	seeing word	15	48	.31	.011
	be guess	9	14	.64	<.001
	cannot remember	41	144	.28	<.001
	not remember	29	89	.33	<.001
	I do	28	99	.28	.004
	at all	11	34	.32	.021
Yes-Guess	think I would	10	21	.48	<.001
	I think I	16	47	.34	.005
	but I cannot	9	25	.36	.026
	I recall	5	10	.50	.015
	but cannot	12	36	.33	.013
	think I	21	77	.27	.020
	but I	23	92	.25	.036
	I think	25	88	.28	.006

Yes-Maybe	think I remember	7	13	.54	<.001
	I think it	7	13	.54	<.001
	when I see	14	23	.61	<.001
	think I will	6	13	.46	.013
	not hundred percent	10	15	.67	<.001
	I see it	14	17	.82	<.001
	presented with it	6	10	.60	<.001
	hundred percent sure	9	13	.69	<.001
	would recognise it	5	11	.45	.025
	word but I	10	28	.36	.018
	I am not	13	41	.32	.018
	to do with	11	19	.58	<.001
	I remember seeing	12	35	.34	.010
	remember other word	7	16	.44	.010
	but not sure	7	15	.47	.006
	word but not	9	18	.50	<.001
	something to do	11	16	.69	<.001
	but I am	9	27	.33	.034
	I think I	17	47	.36	<.001
	I would recognise	7	20	.35	.037
	I can remember	12	33	.36	.008
	second word	8	22	.36	.021
	word and	15	37	.41	<.001
	I know	9	18	.50	.001
	see it	14	18	.78	<.001
	tried to	8	10	.80	<.001
	I will	21	73	.29	.011
	percent sure	9	13	.69	<.001
	I see	18	34	.53	<.001
	exact word	5	11	.45	.025

presented with	10	26	.38	.007
when I	15	35	.43	<.001
other word	15	46	.33	<.001
think I	28	77	.36	<.001
word that	8	22	.36	.021
but I	29	92	.32	<.001
and think	9	10	.90	<.001
it when	6	13	.46	.013
word when	7	13	.54	.002
something to	11	17	.65	<.001
word but	23	72	.32	.001
it was	18	64	.28	.019
but not	20	40	.50	<.001
to do	11	19	.58	<.001
word was	8	23	.35	.042
remember other	7	16	.44	.010
am not	16	47	.34	.005
it I	5	10	.50	.015
not hundred	10	15	.67	<.001
I think	35	88	.40	<.001
hundred percent	12	18	.67	<.001
it is	17	51	.33	.004
do with	11	19	.58	<.001
words I	6	15	.40	.027
will recognise	7	14	.50	.004
think it	8	15	.53	.002
with it	13	32	.41	.001
I remember	41	139	.29	<.001
not sure	18	61	.30	.014
recognise it	13	37	.35	.006

	I feel	10	24	.42	.003
	it but	9	22	.41	.006
Yes-Sure	I remember this	7	16	.44	.010
	so I am	5	12	.42	.036
	I remember word	10	24	.42	<.001
	I thought of	5	10	.50	.015
	I can remember	12	33	.36	.008
	in my head	13	20	.65	<.001
	which is	5	10	.50	.015
	because I	14	31	.45	<.001
	words together	7	15	.47	.007
	my head	13	27	.48	<.001
	two words	15	38	.39	<.001
	I remembered	14	20	.70	<.001
	word association	5	10	.50	.015
	remember this	15	56	.27	.049
	I had	6	15	.40	.027
	remember word	18	68	.26	.049
	remember that	8	16	.50	.002
	I remember	43	139	.31	<.001
	I imagined	6	10	.60	.002
	I thought	9	26	.35	.029
	association between	7	13	.54	.002
	thought of	8	13	.62	<.001
	in my	23	55	.42	<.001
	this is	6	14	.43	.019
	can remember	16	40	.40	<.001
	I can	24	50	.48	<.001
	this pair	5	11	.45	.025
	word in	5	10	.50	.015

I made	11	24	.46	<.001

Note. A count of occurrences of each *n*-gram in justifications for the corresponding JOL category are reported along with total number of occurrences, proportion of occurrence and p-value computed using the binomial test.

As seen in Table 4.4, the types of descriptions for the highest confidence *no* and *yes* responses correspond to 0% (e.g., "do not remember") and 100% (e.g., "I can remember") responses of Experiment 5. It is noteworthy that the high confidence JOLs and *yes* JOL predictions refer to not just the target, but also memory for the "word association" or "link between" the items. This supports recent findings that memory for associations made between the cue and the target at study influences metacognitive confidence (Hertzog et al., 2014) and demonstrates that this is true even when participants are not instructed to use any specific memory techniques in learning the cue-target pairs.

The *guess* responses (for both *yes* and *no* JOL predictions), were fairly low on unique *n*-gram use compared to the other JOLs. The LSA results revealed that participants were more likely to reference the cue than the target for these responses but the *n*-gram results are not clear as to which way this was. However, the *tri*-gram "not remember seeing" occurred 10 times in justifications for the *no-guess* responses (as compared to 12 occurrences for *no – sure* and 9 occurrences for *no – maybe*). While this proportion of occurrence for *no – guess* justifications was only marginal (p = .052), altogether, these results show that references to lack of cue familiarity were reserved for *no* JOL predictions. Consequently, it seems likely that if there is a distinction between *yes* and *no* predictions, it is in whether the cue feels familiar or not.

Nevertheless, the results indicate a less clearly defined distinction between *yes* and *no* responses than some (e.g., Dunlosky et al., 2005) would predict. *Guess* predictions (which here capture low magnitude SOJs) might just be what the term suggests—instances where participants do not feel strongly predisposed toward a *yes* or a *no* prediction and rather the evidence available to them (or its lack) makes them uncertain about the future retrieval status of the items they are evaluating. If anything, this highlights the usefulness of allowing participants to express uncertainty. If one were to interpret the character of the *yes/no* distinction, it is the closest to the differentiation between 0 and 20% JOL.

Lastly, some phrases were almost equally likely for all of the *no* predictions. Namely "I do not", "do not remember", "cannot remember" and "not remember seeing". This indicates that participants were less clear on how to differentiate the three *no* response types from each other and were inclined towards using similar responses across all three confidence levels associated with *no* predictions. Together with the results from Experiment 5, these results suggest that if there is an underlying *yes/no* distinction in the JOL confidence scale, it is likely located at the low-ends of a numeric scale, with the majority of the scale above this point consistent with use of *yes* predictions. This is consistent with framing effects which suggest that participants primarily accrue evidence toward a *yes* prediction as indicated by their judgments being swayed by whether the question is phrased in terms of forgetting or remembering (Finn, 2008; Koriat, Bjork, Sheffer, & Bar, 2004; Serra & England, 2012).

4.3.4 Support Vector Machine (SVM) analysis

Our final analysis was to evaluate the extent to which the written justifications for any two JOL response types were quantifiably distinct. Within each experiment, we trained SVM classifiers to compare each JOL category against all other JOL categories. If two JOL categories were justified by referring to different types of evidence, then classification accuracy for distinguishing the two categories would be good. The results are reported in Table 4.5, which presents overall SVM classifier performance for all JOL categories expressed as percentage of examples classified correctly.

Table 4.5: **Bivariate SVM classification accuracy results by experiment.**

Experiment 5

	20%	40%	60%	80%	100%
0%	75.9	81.7	86.3	93.1	94.7
20%		57.9	73.7	79.9	84.5
40%			59.1	69.9	80.2
60%				60.3	69.0
80%					53.3

50-60
60-70
70-80
80-90
90-100

Experiment 6

	No-Maybe	No-Guess	Yes-Guess	Yes-Maybe	Yes-Sure
No-Sure	76.4	64.2	85.3	91.3	92.2
No-Maybe		62.7	65.7	71.1	91.2
No-Guess			67.5	80.2	90.1
Yes-Guess				62.4	86.4
Yes-Maybe					84.6

Note. The results express percentage of test cases classified accurately and reflect the degree to which two JOL categories could be said to differ in how they were justified.

Examining all adjacent JOL confidence levels, Experiment 5 revealed that the 0% and 20% JOLs were classified with the highest degree of accuracy (this performance was significantly different from the classification performance in the next i.e. 20% vs 40% comparison; X^2 = 9.13, p = .003). This would agree with the proposal that if there is a divergence in processes relied on in making the confidence judgments, it is located between the lowest two points on the scale. All other JOL confidence levels would appear to be graded variations of a similar process (the highest classification accuracy between these of 60.3% was not significantly different from chance performance of 50%; X^2 = 2.31, p = .129).

In Experiment 6, the highest adjacent classification accuracy was between *yes-maybe and yes-sure* predictions, which was significantly higher than the classification accuracy between the *yes* and *no* prediction boundary (i.e. the *guess* responses); X^2 = 11.84, p < .001. This is consistent with the *n*-gram results which showed there were very few distinct features (*bi*-grams and *tri*-grams) characterizing the *guess* responses but contrary to the prediction that *yes* vs. *no* predictions should be highly classifiable (Dunlosky et al., 2005).

Overall, the highest confidence *yes* prediction was well classified in contrast to all other responses (Experiment 6) whereas the 100% responses' classification compared to other high JOL confidence responses (60% and 80% JOL) approached chance performance (50%; Experiment 5). If participants treated most (if not all) of the JOL confidence scale as accumulation of evidence toward a *yes* prediction then it follows that the JOL confidence levels were more clearly defined when there were fewer options provided for a positive prediction. This is in line with other research (e.g., Finn, 2008; Koriat, Bjork, Sheffer, & Bar, 2004) which has shown that participants need to be asked to predict their own forgetting to treat the confidence scale as also expressing the degree to which they might forget (i.e. a *no* prediction) as compared to only the degree to which they might remember (or what we would classify as a *yes* prediction).

In contrast, the *no* responses of Experiment 6 were less clearly demarcated (as compared to the *yes* predictions). As we saw from the *n*-gram analysis, there was a great deal of overlap between the *n*-grams participants used as a way of classifying their *no* predictions. Overall, it seems that in a paradigm where participants aim to predict their remembering, they struggle to differentiate between different levels of not remembering (or forgetting). This is again consistent with the idea that participants would primarily focus on the familiarity of the cue as a way of rejecting future target memory. Cue familiarity is a less varied type of signal than the more heterogeneous nature of different levels and types of target access that would be thought to characterize the unique *yes* JOL predictions.

Most relevant in regards to the current study, the classification pattern for the two response formats is clearly different. This suggests that while there is a distinction in the types of processes driving the JOL confidence responses, it might be troublesome trying to assign them a discrete *no* vs *yes* prediction status. Rather, the two response formats might encourage related but nevertheless different modes of evaluation.

4.4 Discussion

Within any metacognitive paradigm, aspects of the task are manipulated so that specific information is made salient to participants; in metamemory tasks this is usually through encoding or retrieval instructions. The question that arises is whether the information that is shown to influence metacognitive judgments in such paradigms remains relevant in other contexts (see for example Hertzog et al., 2014). This study asked: what information do participants consider relevant to their JOLs in the absence of any such manipulation and how does this information map onto theory? More specifically, we investigated spontaneous written justifications for numeric confidence and binary (*yes/no*) JOL predictions. Participants completed a standard JOL task and on some trials were asked to justify their predictions, which were subsequently analysed using a range of natural language processing techniques. The results showed that (i) participants could justify their metacognitive judgments, (ii) confidence JOL justifications mapped broadly onto current theory as they referenced both cue and target related information, (iii) confidence JOLs had different characteristics to binary JOLs.

Overall, participants were able to justify their JOLs and did so with reference to both cue- and target-related information as well as with reference to associations they made between them. This was even though we did not manipulate these factors nor did we instruct participants in any way as to how they should learn the items and what information they should focus on when making their JOLs. The results thus complement studies which have shown that emphasis on cue, target and associative information shifts metacognitive confidence (Benjamin, 2005; Hertzog et al., 2014; Metcalfe & Finn, 2008b) and support the heuristics view of metacognitive judgments as based on evidence accumulation processes (Brewer, Marsh, Clark-Foos, & Meeks, 2010; Koriat, 2000).

The results of confidence JOLs were consistent with the predictions of Metcalfe and Finn (2008b). The 0% and 20% JOL responses were the most divergent of any adjacent JOL confidence levels as indicated by highest classification accuracy. The content analyses supported the idea that, whereas the 0% JOLs corresponded to a lack of cue familiarity, the 20% JOLs were given to items whose cue was familiar but whose target was not accessible. All other JOL confidence levels reflected an increase in target accessibility. We therefore provide support for an account of JOL confidence as resulting from a two-stage evaluation, with interrogation of different evidence characterizing each stage.

The results of Experiment 6 suggested that participants referred to the cue to distinguish between a *no* and some degree of a *yes* response as well as to characterize high confidence *no* responses from all other responses. This would map onto the differences between 0% and 20%, suggesting that if there is an underlying *yes/no* distinction in the JOL confidence scale, it is a differentiation of the lowest confidence responses only. Consistent with this, there is also an indication that participants struggled to distinguish three different levels of *no* confidence predictions from each other, at least when framed in terms of remembering. The degrees of *yes* predictions were more clearly demarcated from each other. However, the overarching distinction between *yes* and *no* predictions was less clear-cut than predicted (e.g., Dunlosky et al., 2005) and it remains questionable whether *yes* vs. *no* responding reflects how participants approach the JOL confidence scale.

At the very least, it is clear that analogous points on the two 6-point scales were not equivalent—we cannot treat the numerical JOL confidence scale as evenly split into *yes* and *no* responses. Across the two question formats, both ends of the scale corresponded, i.e. a 0% JOL was equivalent to a high confidence *no* and a 100% JOL was equivalent to a high confidence *yes*. It is unsurprising that our understanding of the extremes of the scale might be correct. However these extremes differed in how they related to the mid-range responses and this is where we observed the most differences. The overall different pattern of JOL justifications across the two response formats highlights that participants do not use all points of the two scales in the same ways.

This lack of equivalence is worth highlighting, especially as there is an underlying assumption in much metacognitive research that confidence judgments are probabilistic. It is common, for example, to interpret 0%, 20% and 40% as *no* predictions. This is seen particularly in assessments of metacognitive accuracy in terms of calibration; an assessment of whether metacognitive judgments correspond exactly to performance (perfect calibration would be for items given 60% JOLs were recognized at a rate of 60% in subsequent memory tests etc.). Considerable research has gone into understanding what drives poor calibration which is observed across domains (see for example Finn & Metcalfe, 2007; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Koriat, Lichtenstein, & Fischhoff, 1980; Koriat, Ma'ayan, Sheffer, & Bjork, 2006; Kornell & Bjork, 2009). However, recently Hanczakowski et al. (2013; see also Zawadzka & Higham, 2015) showed that the common observation that participants tend to display underconfidence in terms of calibration (i.e. lower average confidence JOL than overall memory performance) is not observed with a *yes/no* response format and when the proportion of *yes* responses is used to assess calibration. Hanczakowski et al. interpreted this as indicating that participants are not truly underconfident as has been previously assumed and that the results could rather be explained as driven by misunderstanding of how participants treat the JOL confidence scale. This finding is consistent with the suggestion from the current study that participants are treating most of the JOL confidence scale as a *yes* prediction.

This further relates to findings that question format influences how participants respond in both metacognitive (Finn, 2008; Serra & England, 2012) and recognition memory tasks (Mill & O'Connor, 2014). For example, participants anchored their JOLs lower on the JOL confidence scale when judging future remembering as compared to forgetting (Serra & England, 2012). Similarly, recognition judgments for whether an item has been studied or is seen for the first time have been shown to be influenced by whether the question is termed in terms of judging 'oldness' or 'novelty' (Mill & O'Connor, 2014). More specifically, participants shifted their response bias to more likely disconfirm the question asked (more likely to respond 'new' when asked 'old?'). This study adds to a newly growing literature demonstrating that, in addition to question format, response format also influences participant responding in metacognitive tasks (Jersakova, Moulin, & O'Connor, 2016; Overgaard & Sandberg, 2012). Taken together, these studies demonstrate that consideration of the methods used to assess a cognitive or a metacognitive phenomenon is of theoretical importance, with direct consequences for the inferences we draw from our data. In processes that are characterized as evidence evaluation, it is remarkable that the question and response format can influence what type of evidence participants consider and how they interpret it. Such findings help to further elaborate on the evidence evaluation processes that underlie participant responses in cognitive tasks.

A further question of interest is how the present results might be of relevance to other metamemory and metacognitive tasks. A self-apparent comparison is to the immediate JOL paradigm, in which participants render judgment immediately after study, with both the cue and the target present. Immediate JOLs have been demonstrated to differ from delayed JOLs and to rely on different types of evidence (Koriat & Bjork, 2006; Rhodes & Tauber, 2011). More specifically, whereas delayed JOLs require an evaluation of access to information in long-term memory, immediate JOLs rely primarily on information held in short-term memory. Consequently, one would expect to observe different patterns of responses and distinct content in justifications for immediate as compared to delayed JOLs; for example, participants do not need to attempt to retrieve the target which is present in immediate JOLs and they might instead focus on the level of association between the cue and the target (Koriat & Bjork, 2005).

The expectation isn't that the extension of the present paradigm to other metacognitive tasks would produce exactly the same results. Based on the current findings, our prediction would be that participants would be able to produce justifications for their immediate JOLs (and other metacognitive judgments) and that these would reference pertinent, task specific information (e.g. the relationship between the cue and the target). A similar paradigm employed in other types of metacognitive tasks is likely to confirm that they can collectively be considered evidence aggregation and evaluation processes. Conversely, it would be of interest to investigate which types of influences participants might not be aware of through failing to account for them in their justifications. Equally, it is possible that under particular circumstances the confidence scale can be interpreted as equally split into binary *yes/no* responses even though it does not seem to be the case with delayed JOLs. The key message of this chapter is that this should not be assumed before it is confirmed for the given experimental context under which the particular metacognitive paradigm is being investigated.

In summary, we provide evidence for metacognitive confidence judgements as resulting from evaluative processes that weigh the degree of evidence toward the decision framed by the response-eliciting question (in this case, ‘will this item be remembered?’). The present study demonstrates that participants have at least a degree of access into this process and can justify the JOLs they are making. What is more, they do so with reference to processes observed to influence JOL magnitude in the literature, primarily memory related signals relating to the cue (how familiar it is) and the target (how about associated information can be accessed). Importantly, the results demonstrate that widely used numeric confidence JOLs are unlikely to have an underlying, direct *yes/no* mapping. At the very least, this distinction is unlikely to be couched in probabilistic terms (e.g., 40% interpreted as a rejection of future retrieval). This finding should guide future interpretations of metacognitive confidence judgments and encourage researchers to avoid making unwarranted assumptions about what participants’ confidence judgments represent. The particular anchoring of the confidence scale is likely to shift between tasks and participants, and the distance between the points on the scale (in terms of the strength of evidence they refer to) is also subject to shifts (Zawadzka & Higham, 2016). Further, these results highlight the need for confirming results across response formats.

CHAPTER 5

THE COMPARATIVE AND RELATIVE NATURE OF DELAYED JOLs

5.1 Introduction

In his cue-utilization framework, Koriat (1997) suggested that there are a number of cues (intrinsic, extrinsic and mnemonic) that influence JOLs (see section 1.3.3 of the Introduction for more detail). Briefly, intrinsic cues refer to characteristics of the studied items (e.g., level of relatedness between the cue and the target), extrinsic cues refer to the entire learning situation (e.g., time given to study each item), and mnemonic cues refer to internal and subjective experiences (e.g., feeling of knowing the sought after target information). All the preceding chapters have examined the role of stimulus specific (mnemonic) cues (e.g., cue familiarity) in delayed JOLs. In contrast, this chapter looks at the effects of extrinsic, theory-based influences. In keeping with the central theme of this thesis, we examined whether delayed JOLs are sensitive to the memory test they are predicting (recognition as compared to recall). The major focus of this chapter is on whether this is dependent on the test and response format employed.

One of the oldest and most reliable findings in memory research is that recognition memory is superior to recall (MacDougall, 1904). Although there are exceptions to this rule when one considers extralist or nonlist cues (i.e. not directly studied but related items), in a standard cue-target learning paradigm where participants are presented with the studied cues at test and either asked to recall the target or to choose it from a list of options, participants perform better in the latter task (Tulving & Thomson, 1973). Consistent with this, past research has shown that participants are sensitive to memory test format (i.e. recognition/recall) when making delayed JOLs if the memory test predicted is manipulated on a trial-level, in a within subject design or if they have past experience with the memory test format (e.g., multiple study-JOL-test cycles; Mazzoni & Cornoldi, 1993; Thiede, 1996). This could seem to suggest that participants are sensitive to theories about memory when making delayed JOLs. However, it is also possible that the results are specific to the experimental design employed.

The immediate JOL literature provides a good illustration of this idea. It has been demonstrated that immediate JOLs are also sensitive to extrinsic, theory based influences but only when these are made salient to participants i.e. in trial-level as compared to blocked designs (Koriat, Bjork, Sheffer, & Bar, 2004; Kornell & Bjork, 2009). For example, when participants were asked to predict their ability to remember recently learned information in two weeks time as compared to the following day or the following year, participants only changed their JOLs with retention interval when the test delay predicted changed on a trial level as compared to a blocked or between subject design (i.e. participants would on each trial predict retention for a different test delay rather than only one delay throughout the entire task). This was even though they clearly believed that forgetting increases over time when asked independently of the task. This has been termed the stability bias and interpreted as indicating that participants assume their memories remain relatively stable over time when making online, trial-level judgments (Koriat et al., 2004; Kornell & Bjork, 2009).

Alternatively, this is consistent with Koriat's (1997) observation that immediate JOLs are comparative in nature. More specifically, if the study list is composed of different sets of items (e.g., related as compared to unrelated pairs of words), participants will use this contrast between the items when constructing their JOLs (Castel et al., 2007; Tiede & Leboe, 2009). Similarly, metacognitive illusions, such as JOLs changing with perceptual fluency of the studied items (manipulated by font-size; Rhodes & Castel, 2008), only hold in mixed-list and not in between-subject designs (Susser, Mulligan, & Besken, 2013).

This further relates to recent findings that confidence judgments given in immediate JOLs are relative in nature rather than absolute (Hanczakowski et al., 2013; Zawadzka & Higham, 2015). As mentioned previously, these studies argued that participants do not use confidence in probabilistic terms and do not attempt to ascertain the probability of future retrieval or the percentage of items they will retrieve. Instead, participants may use confidence to rank the items against each other with high confidence corresponding to items most likely to be retrieved on the subsequent memory test in comparison to other items on the list. As such confidence judgments on any given trial are made in relation to other judgments made and stimuli encountered on the same task (see Rahnev et al., 2015 for similar arguments about perceptual metacognitive judgments). It is likely that these findings are not particular to immediate JOLs.

We wanted to explore whether delayed JOLs could also be characterized as comparative and relative. We did this in the context of exploring whether participants are sensitive to the memory test they are predicting by investigating whether test format modulates this effect. If participants were sensitive to the memory test (recognition vs. recall) they were predicting irrespective of the test format employed (trial-level vs. blocked), one could conclude that delayed JOLs are actually sensitive to theory-based (explicit) influences. If this effect were to be modulated by test format, this would indicate that delayed JOLs are instead comparative and relative in nature. As such, across two experiments, we manipulated whether participants predicted future recognition or recall and whether they made both predictions in an intermixed, trial-level design (Experiment 7a) or in a blocked design (Experiment 7b).

Another variable of interest was whether results would generalize across confidence and binary JOL responding. Chapter 4 demonstrated that the two response formats do not necessarily map directly onto each other. This adds to a growing literature that highlights the need to confirm patterns of results across multiple judgment response formats (Hanczakowski et al., 2013; Jersakova et al., 2016). This is to ensure that the observed results are not simply driven by the response format employed. In this chapter we turn to the direct, within-subject comparison of confidence and binary JOLs and the implications this has for study of delayed JOLs.

In summary, we manipulated (a) whether participants predicted recognition or recall with half of trials assigned to each memory prediction, (b) whether these predictions were made side-by-side in a mixed design or whether they were blocked with only one prediction made at a time and (c) whether participants gave confidence or binary JOLs (again, one judgment response format employed for half the trials). Whereas (a) and (c) were manipulated within subject in a 2x2 design, (b) was manipulated between subject across the two experiments. We examined whether participants would change their JOLs according to whether they were predicting future recognition or recall, whether this would be dependent on the design and whether the results would generalize across JOL response formats. Altogether, this allowed us to explore whether (i) delayed JOLs are relative in nature and (ii) how confidence and binary JOLs compare. We expected that participants would perform significantly better on the recognition as compared to the recall task. Based on results from the immediate JOL literature (e.g., Susser et al., 2013), we anticipated that participants would only be sensitive to this difference in the memory test performance they were predicting in the trial-level design. Similarly, due to other findings in the immediate JOL literature (Hanczakowski et al., 2013), we also expected to see differences in calibration and overall JOL responding between the different JOL response formats but not in resolution accuracy.

5.2 Method

5.2.1 Participants

All participants were native English speakers affiliated with the University of Leeds (students and staff) with 25 participants (4 men, mean age = 23.0, *SD* = 5.6) in Experiment 7a and 26 participants (7 men, mean age = 26.3, *SD* = 9.6) in Experiment 7b. Participants either took part for course credit or were reimbursed £3 for taking part. They were only allowed to participate in one of the experiments. The study was granted ethical approval by the School of Psychology, University of Leeds ethics review board.

5.2.2 Materials

For each participant, the studied items were randomly selected from a list of 628 common, singular English nouns (5-6 letters long) taken from the English Lexicon Project (minimum log Hyperspace Analogue to Language frequency 8.02; Balota et al., 2007). Each participant was exposed to a unique set of 60 cue-target word-pairs (e.g., truth-eagle).

5.2.3 Procedure

Across two experiments, we used a Judgment format (binary, scale) x Memory test predicted (recognition, recall) within subjects design. This means that all participants completed all conditions. The difference between Experiment 7a and Experiment 7b was in how the Memory test manipulation was employed (in a trial-level vs. blocked design). See Figure 5.1 for a schematic of the procedure in Experiment 7a.

Participants were not informed at the beginning of the study as to how their memory will be tested, rather, they were only told to memorize the pairs. A deviation from previous chapters was an introduction of study instructions, making sure participants deeply encoded all pairs. Due to the experimental design we needed to expand the number of cue-target pairs but the core questions of this chapter prevented us from using multiple study-test cycles as was done in previous chapters. This was because experience with the test-format would influence participant responses on subsequent cycles (see Thiede, 1996). To ensure that participants were not performing at floor in recall, participants were instructed to use the most effective associative learning strategy: mental imagery (Bower, 1970; Bower & Winzenz, 1970).

The general procedure was the same as in previous chapters with subsequent Study, Judgment and Memory Test Phases (see Figure 5.1). In the Study Phase, participants first learned 60 individually presented cue-target word pairs. They were instructed to use mental imagery to remember each pair. This was followed by a Judgment-of-Learning (JOL) Phase where, on presentation of each studied cue participants were asked to predict whether (i) they will recognize the associated target or (ii) whether they will recall it. For half of the trials participants made a binary (*yes/no*) prediction whereas on the other half of the trials participants gave a confidence judgment (0-20-40-60-80-100%) by pressing on the corresponding key. The last part of the experiment was a memory test where, on presentation of each cue, participants were asked to (a) select the associated target from 3 options (all targets from the study) or (b) recall the target and type the answer on the keyboard. After memory for all items was tested consistent with the retrieval test the JOL predicted; participants also completed a recognition test for all items they attempted to recall.

In Experiment 7a, the Judgment Phase was split into two blocks (counterbalanced across participants). Each block started with instructions informing participants of the judgment response format they were to use in that block. In one block participants made binary (*yes/no*) JOLs while in the other block they gave confidence responses expressed as a percentage (0%-20%-40%-60%-80%-100%). The retrieval format (whether recognition or recall was predicted) was manipulated on a trial level within-block. Participants were presented on screen with the cue-word as well as an instruction word at the top of the screen (*Recognise?* or *Recall?*). It was ensured that half the trials within each JOL format block received each type of prediction. The order in which the judgments were made within each block was randomised.

In Experiment 7b, the Judgment Phase was split into four blocks (counterbalanced across participants). Again, each block started with a set of instructions explaining the JOL response format to use *and* the memory test participants were to predict. Across the four blocks participants were asked to (a) predict whether they will recognize the target using a binary response format, (b) predict whether they will recognize the target using a confidence scale, (c) predict whether they will recall the target using a binary response format and (d) predict whether they will recall the target using a confidence scale. Same as in Experiment 7a, the cues were accompanied by an instruction word (*Recognise?* or *Recall?*) at the top of the screen to ensure that participants remembered which prediction they were making in each block.

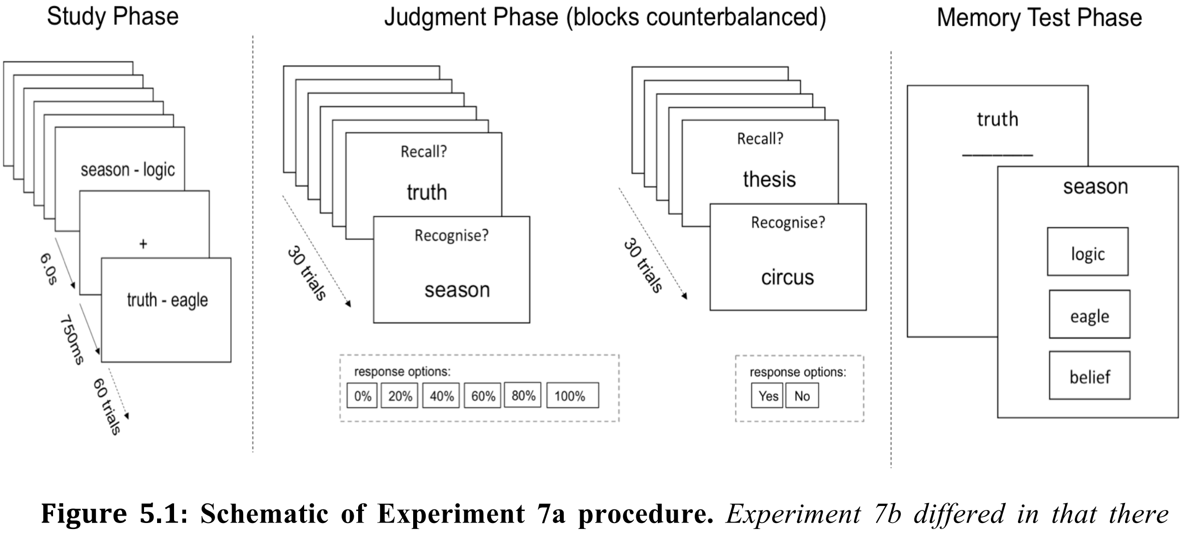

Figure 5.1: Schematic of Experiment 7a procedure. *Experiment 7b differed in that there were 4 blocks in the Judgment Phase and participants only predicted recognition or recall within each block (never both).*

5.3 Results

5.3.1 Memory

Firstly, we compared average memory performance (see Figure 5.2) between conditions and experiments using an Experiment (7a, 7b) x Judgment format (binary, confidence) x Memory test (recall, recognition) ANOVA. Average recognition performance was significantly higher than recall performance, $F(1, 49) = 514.62$, $p < .001$, $\eta_p^2 = .91$. There was no difference in memory performance between whether confidence or binary JOLs were used, $F < 1$, and between experiments, $F < 1$.

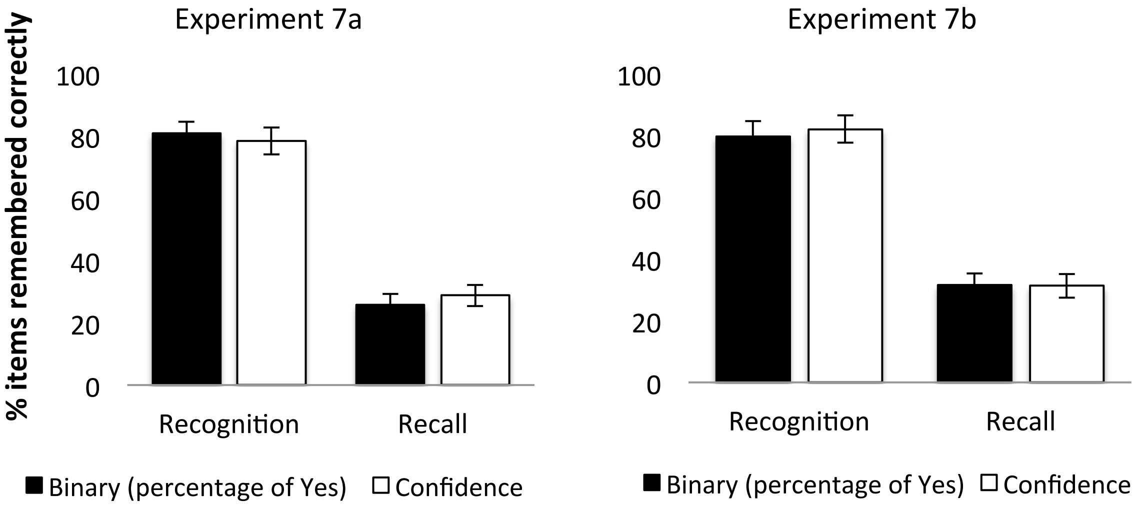

Figure 5.2: Mean percentage of items correctly remembered (recall and recognition) in Experiment 7a and 7b by JOL response format. *Error bars indicate standard error of the mean.*

5.3.2 JOL responding

In line with previous studies (e.g., Kornell & Bjork, 2009), we first looked at absolute correspondence between number of items remembered and the average JOL expressed. One of the key variables of interest was whether JOL predictions would be sensitive to the retrieval test they were predicting and whether this would change with (i) the nature of the experimental design (blocked vs. trial-level) and (ii) the JOL response format employed (confidence vs. binary JOL). To investigate this we compared the average JOL confidence (expressed as percentage) on recognition and recall trials with the average percentage of positive (*yes*) predictions of binary JOLs given. The fact that both yielded percentages allowed us to compare the two JOL response formats directly (see Figure 5.3).

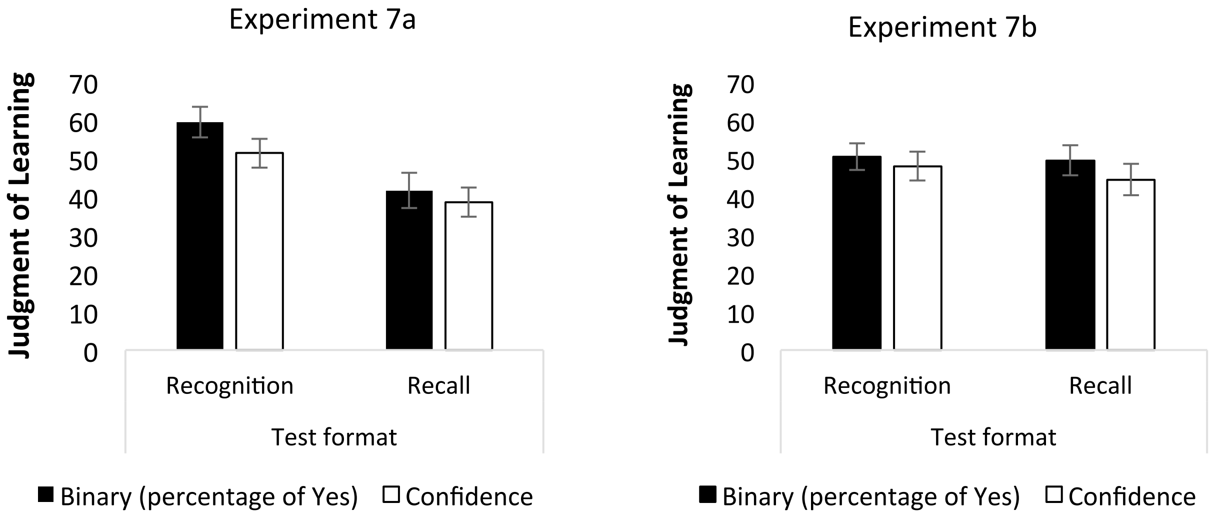

Figure 5.3*:* Average JOLs for recognition and recall predictions expressed as a binary judgment (percentage of y*es* predictions) or overall confidence expressed for Experiment 7a and Experiment 7b. *Error bars represent standard error of the mean.*

We carried out the same Experiment (7a, 7b) x Judgment format (binary, confidence) x Memory test (recall, recognition) ANOVA as before, this time to analyse mean JOL responding. While there was no main effect of Experiment, $F < 1$, there was a main effect of both Memory test, $F(1, 49) = 16.91$, $p < .001$, $\eta_p^2 = .26$, and Judgment format, $F(1, 49) = 314.83$, $p < .001$, $\eta_p^2 = .86$. Further, there was also a two-way interaction between Memory test and Experiment, $F(1, 49) = 5.60$, $p = .022$, $\eta_p^2 = .10$, and an interaction between Judgment format and Memory test, $F(1, 49) = 16.35$, $p <.001$, $\eta_p^2 = .25$, but not between Judgment format and Experiment, $F < 1$. Lastly, the results also showed a three-way interaction between all factors of interest, $F(1, 49) = 5.28$, $p = .026$, $\eta_p^2 = .10$.

To understand the interactions more closely, we split the data by experiment and analysed these separately using a Judgment format x Memory test ANOVA. We found that in Experiment 7a, higher JOLs were given for recognition as compared to recall predictions, $F(1, 24) = 31.97$, $p < .001$, $\eta_p^2 = .57$. There was no effect of Judgment format, $F(1, 24) = 2.83$, $p = .105$, $\eta_p^2 = .11$, or the interaction, $F(1, 24) = 1.21$, $p = .283$, $\eta_p^2 = .05$. In Experiment 7b on the other hand, there were no differences between recognition and recall predictions in the magnitude of JOL expressed, $F < 1$. Again, there was no effect of Judgement format, $F(1, 25) = 1.56$, $p = .223$, $\eta_p^2 = .06$, or the interaction, $F < 1$.

In summary, the above results show that JOLs were higher for recognition than recall when the type of retrieval predicted changed on a trial level (Experiment 7a) but not when memory test predicted (recall, recognition) was blocked (Experiment 7b). This pattern of results did not differ with the type of JOL response format (binary, confidence scale) used. However, overall participants seem to have given a higher proportion of *yes* responses than the average JOL confidence that they expressed (as indicated by the main effect of Judgment format). The three-way interaction between all factors of interest further shows that this was not consistently so for all conditions. For example, in Experiment 7a the magnitude of difference between the two response formats is bigger for recognition than recall predictions. In experiment 7b, it is the opposite. As such there seem to be small variations in how the JOL response formats are employed even though the key main effects of interest and interactions remain the same.

5.3.3. Absolute accuracy (calibration)

The above analysis shows that in Experiment 7a JOL responding differed between recognition and recall predictions. Nevertheless, it is not clear how close these predictions were to actual memory performance. One way to assess JOL accuracy is in calibration terms – comparing whether the average JOL percentage (confidence or yes responding) is the same as the percentage of items correctly remembered. For recognition predictions we looked at recognition accuracy and for recall predictions we looked at recall accuracy. Within each experiment, we conducted a Measure (memory performance, JOL) x Memory test (recall, recognition) x Judgment format (confidence, binary) ANOVA.

In Experiment 7a, there was a main effect of Measure indicating that JOL responses and memory performance differed, $F(1, 24) = 8.12$, $p <.01$, $\eta_p^2 = .25$. There was no effect of Judgment format, $F(1, 24) = 1.57$, $p = .222$, $\eta_p^2 = .06$, but a main effect of memory test, $F(1, 24) = 151.64$, $p <.001$, $\eta_p^2 = .86$. There was also a Measure x Memory test interaction, $F(1, 24) = 220.70$, $p < .001$, $\eta_p^2 = .91$, with no other significant interactions (lowest p value = .104). Even though JOLs changed with memory test, participants overestimated recall, $t(24) = 5.30$, $p < .001$, $d = 0.67$, and underestimated recognition, $t(24) = 10.45$, $p <. 001$, $d = 1.43$.

In Experiment 7b, unsurprisingly, there was a main effect of Measure as average JOL responses and memory performance differed, $F(1, 25) = 6.89$, $p <.05$, $\eta_p^2 = .22$. There was no main effect of Judgment format, $F < 1$, but a main effect of Memory test, $F(1, 25) = 161.75$, $p < .001$, $\eta_p^2 = .87$. This was further qualified by a Measure x Memory test interaction, $F(1, 25) = 163.47$, $p < .001$, $\eta_p^2 = .87$, which is consistent with results reported above showing that whereas in terms of memory performance, recognition and recall differed, JOL responses were the same for both memory test predictions. Same as in Experiment 7a, recall was overestimated, $t(25) = 4.28$, $p < .001$, $d = 0.77$, and recognition was underestimated, $t(25) = 9.05$, $p < .001$, $d = 1.79$. There were no further interactions reaching significance (lowest p value = .103).

The calibration analysis shows that in both experiments, participants failed to fully capture the difference in memory performance between recognition and recall. This was true even in Experiment 7a where participants changed their JOL responses based on what they were predicting (unlike in Experiment 7b). The calibration results show that despite the change in responding, participants continued to underestimate recognition and overestimate recall. Whereas calibration has been criticised as an imperfect tool for assessing confidence JOL accuracy unless also generalized to other JOL response formats (Hanczakowski et al., 2013), it is clear that this concurrent over/underestimation of memory performance was equally true for binary and confidence JOL responses.

5.3.4 Relative accuracy (resolution)

We also conducted an analysis of relative accuracy. Compared to absolute correspondence between responding and memory performance as assessed above, relative accuracy aims to capture item-by-item correspondence between JOL predictions and memory. We computed *d'* for binary responses (see Figure 5.4) and AUC for confidence responses (see Figure 5.5).

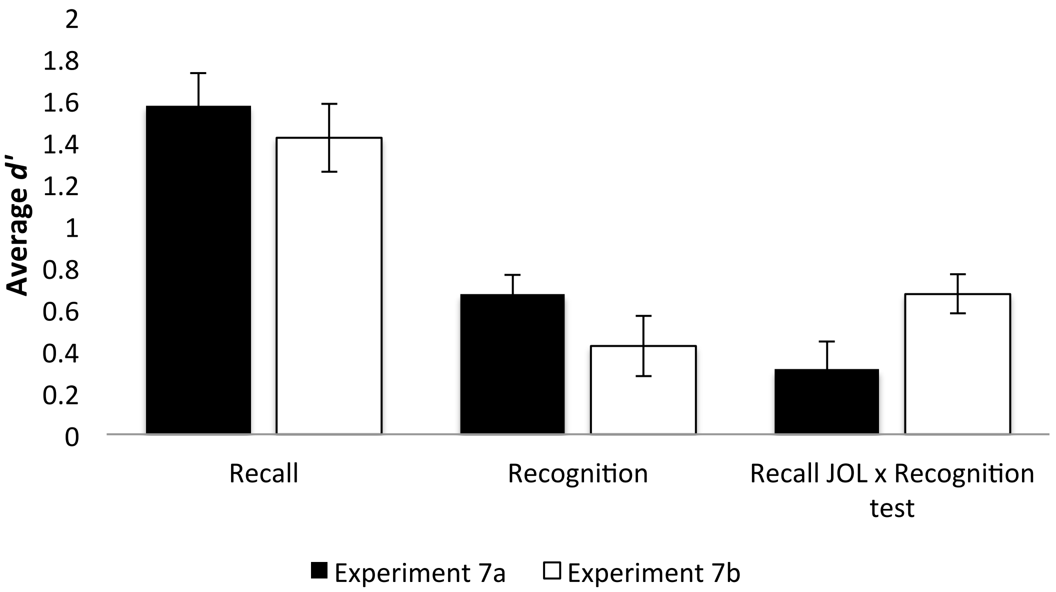

Figure 5.4: Average resolution accuracy for binary responses (*d'*) by Experiment and by Memory test. *Accuracy was first assessed consistent with what participants predicted i.e. if JOLs predicted recall we assessed how accurately participants predicted recall performance and the same for recognition. We also assessed how accurately participants predicted recognition when they made recall JOL predictions. Error bars represent standard error of the mean.*

Firstly, an Experiment (7a, 7b) x Memory test (recall, recognition) ANOVA was used to investigate *d'*. We computed *d'* consistent with JOL predictions; if participants were asked to predict recognition then recognition memory performance was used to determine the hit rate and false alarm rate of the JOL predictions. We observed an effect of Memory test, $F(1, 49) = 57.98$, $p < .001$, $\eta_p^2 = .54$, as recall predictions were more accurate than recognition predictions across both experiments. There was no effect of Experiment, $F(1, 49) = 1.63$, $p = .208$, $\eta_p^2 = .03$, and no interaction, $F < 1$.

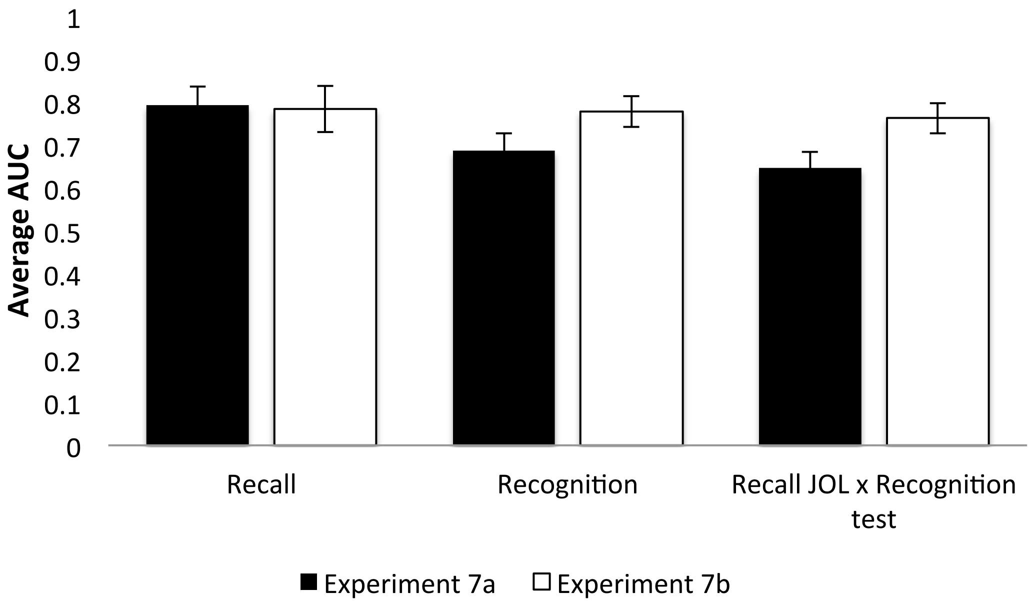

Figure 5.5: Average resolution accuracy for confidence responses (AUC) by Experiment and by Memory test. *Accuracy was first assessed consistent with what participants predicted i.e. if JOLs predicted recall we assessed how accurately participants predicted recall performance and the same for recognition. We also assessed how accurately participants predicted recognition when they made recall JOL predictions. Error bars represent standard error of the mean.*

The same ANOVA was used to analyse the relative accuracy of confidence JOLs (AUC) which showed a different pattern of results. This time there was no effect of Memory test, $F(1, 49) = 1.62, p = .209, \eta_p^2 = .03$, as well as no effect of Experiment, $F < 1$. There was also no interaction between the two factors, $F(1, 49) = 1.31$, $p = .259$, $\eta_p^2 = .03$. All in all, confidence accuracy was the same across Experiments and Memory test predicted.[2] In summary, relative accuracy appears to have differed across the two response formats.[3]

[2] Even though there appears to be a numerical difference in the recall and recognition JOL accuracy in Experiment 7a, this is not significantly different, consistent with the results of the ANOVA (*p* = .068).

[3] We also calculated AUC for the binary data to make sure the difference in the pattern of results was not due to differences in the nature of the tests employed to analyze them. The binary AUC data results showed the exact same pattern as that observed with *d'* i.e. higher recall as compared to recognition accuracy across experiments.

Lastly, we also checked JOL prediction accuracy by whether the target was subsequently recognised or not, irrespective of what the JOL predicted (recall or recognition).[4] This allowed us to compare JOL accuracy when participants predicted recognition as compared to when they predicted recall, keeping the memory performance constant. If participants did not truly differentiate between the two test formats in their predictions, then we would expect the accuracy prediction to be the same. If participants were sensitive to test format then JOL accuracy for recognition predictions should be higher than JOL accuracy when predicting recall if both judgments are assessed in terms of recognition performance. Same as before, we compared *d'* for binary judgments and AUC for confidence judgments (see Figures 5.4 and 5.5).

To analyse each JOL response format and the corresponding resolution accuracy (*d',* AUC) we carried out a JOL prediction (recall, recognition) x Experiment (7a, 7b) ANOVA. The analysis of *d'* showed no main effect of what JOL aimed to predict, $F < 1$, or Experiment, $F(1, 49) = 1.39$, $p = .244$, $\eta_p^2 = .03$, but we did observe a significant interaction, $F(1, 49) = 13.08$, $p < .001$, $\eta_p^2 = .21$. Follow up analyses showed that in Experiment 7a, when participants predicted recognition performance they were more accurate in predicting recognition than when they were predicting future recall performance, $t(24) = 2.62$, $p = .015$, $d = 0.62$. In Experiment 7b, there was no significant difference in how accurately participants predicted their future recognition performance between recognition and recall JOL predictions, $t < 1$.

[4] The primary aim of the experiment was to assess accuracy of JOL predictions consistent with the memory test predicted. To this end we first tested participants' memory consistently with what they predicted (recognition for recognition predictions and recall for recall predictions). This meant that while it was possible to test recognition for items that were previously recalled at the end of the experiment, it was not sensible to test recall for items that were previously tested on recognition. These items were more likely to be remembered since participants saw the targets multiple times. Consequently, we were not able to carry out this specific analysis for the reverse situation as well (recall performance when participants predicted recognition).

However, the same analysis of the AUC data did not yield the same results. There was no effect of what the JOL aimed to predict, $F < 1$, but an effect of Experiment, $F(1, 49) = 5.83$, $p = .020$, $\eta_p^2 = .11$. There was no interaction, $F < 1$. Participants were more accurate at predicting recognition performance (irrespective of what they aimed to predict) in Experiment 7b as compared to Experiment 7a.

5.3.5 Response bias

We also analysed response bias for binary responses (see Table 5.1). This is the tendency to respond either *yes* or *no*, independent of metamemory accuracy. The fact that discrimination (or relative accuracy) is controlled for means that bias (or criterion) is a more sensitive measure of responding effects than looking at the average of positive (*yes*) JOL responses as done earlier.

Table 5.1: **Average criterion by Experiment and memory task.** *Standard deviations appear in parentheses.*

	Recall	Recognition
Experiment 7a	-.05 (*.46*)	-.03 (*.41*)
Experiment 7b	-.15 (*.44*)	.09 (*.45*)

An Experiment (7a, 7b) x Memory test (recognition, recall) ANOVA was used to analyse the criterion results from binary data. There was no effect of memory test, $F(1, 49) = 3.18$, $p = .081$, $\eta_p^2 = .06$, no effect of experiment, $F < 1$, and no interaction, $F(1, 49) = 2.45$, $p = .124$, $\eta_p^2 = .05$. Even though the interaction was not significant, we conducted follow up within Experiment *t*-tests as we were primarily interested in whether participants set the same or different criteria for the recall and recognition predictions within experiment. While there were no differences in Experiment 7a, $t < 1$, participants used a more conservative response criterion for recognition as compared to recall predictions in Experiment 7b, $t(25) = 2.26$, $p = .033$, $d = 0.54$.

This reflects the fact that in Experiment 7b participants did not change their responding based on what they were predicting (recognition or recall). In both tasks, the average percentage of *yes* predictions was about 50% and yet average recall performance was 32% and average recognition performance was 80%. What this means is that recognition predictions would have contained a lot of misses (instances where participants did recognise the item but predicted they wouldn't). Conversely, recall predictions must have contained a large number of false alarms (instances where participants predicted they will remember the item but didn't). In other words, recognition predictions contained fewer *yes* predictions than was warranted given the task (i.e. conservative responding) whereas recall predictions contained more *yes* responses than was warranted (i.e. liberal responding).

For the confidence data, we computed a response criterion for all possible confidence thresholds for a *yes* prediction (as done to compute AUC). The analysis focused on within experiment differences between response criterion placements for recognition and recall predictions (subtracting recall *c* from recognition *c*; see Figure 5.6). A *yes* response threshold (20%, 40%, 60%, 80%, 100%) x Experiment (7a, 7b) ANOVA revealed a main effect of threshold, $F(4, 196) = 3.01$, $p = .019$, $\eta_p^2 = .06$, experiment, $F(4, 196) = 7.52$, $p = .008$, $\eta_p^2 = .13$, and a significant interaction, $F(4, 196) = 2.73$, $p = .031$, $\eta_p^2 = .05$. To better understand the interaction and as we were primarily interested in within experiment criterion differences, within each experiment and for each response threshold, we looked at whether the criterion difference for recognition and recall predictions was significantly different from 0. Because we were analysing five response thresholds, we adjusted p values for multiple comparisons using the Bonferroni correction. In Experiment 7a, there were no significant differences between recall and recognition predictions (lowest p = .070). As such it seems that the criterion placement for recognition and recall predictions was mostly the same across response thresholds in Experiment 7a. In Experiment 7b on the other hand, participants employed a more liberal criterion for recall predictions when using the 20% (t(25) = 2.93, p = .035) and 40% (t(25) = 3.22, p = .020) thresholds with no further significant differences (lowest p = .200). In other words, at least for the lower end of the confidence scale, participants were more conservative in their recognition as compared to their recall predictions. Similar to the results from the binary data, this could account for the lack of a difference in average JOL responding.

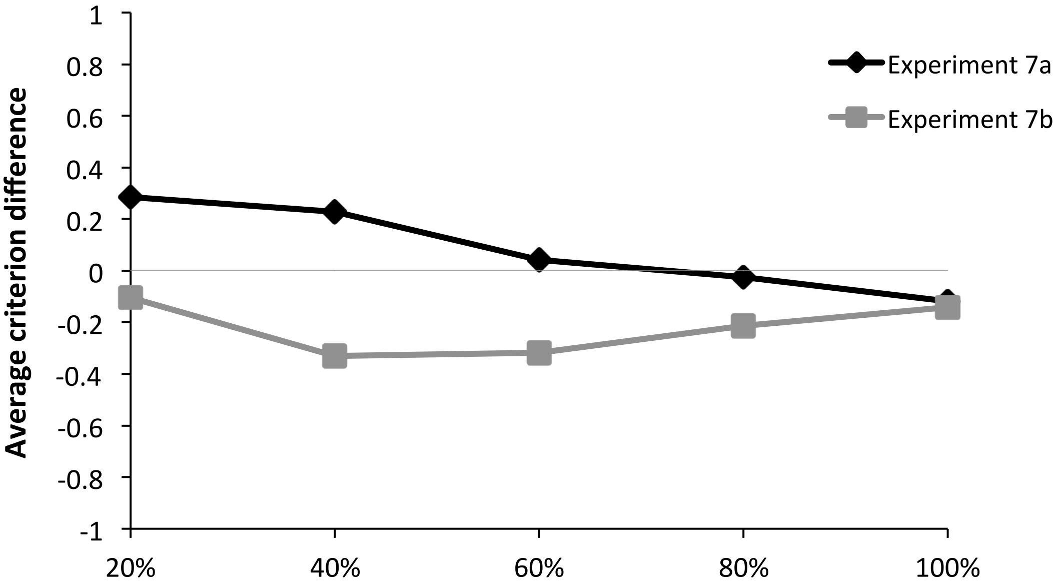

Figure 5.6: Average difference between recall and recognition criterion for each response threshold by Experiment. *Values above zero indicate a more conservative criterion for recall predictions as compared to recognition predictions and values below zero indicate more liberal criterion for recall predictions.*

5.4 Discussion

This chapter examined whether delayed JOLs could be characterized as relative in nature and/or whether they are sensitive to theories about memory. While such questions have been examined with immediate JOLs, they have not been extended to delayed JOLs, which are thought to rely on distinct processes (Koriat, 1997). Therefore, in the present study, on half the trials participants predicted whether they would recognise the target, and on the other half they predicted whether they would recall it. Whereas in Experiment 7a these judgments were inter-mixed, encouraging participants to draw direct comparisons, in Experiment 7b they were blocked. Lastly, participants gave JOLs as either a confidence or a binary (*yes/no*) judgment. This allowed us to evaluate whether results generalize across JOL response formats.

As predicted, recognition memory performance was significantly better compared to recall (MacDougall, 1904). Even though the reasons for this difference remain open to debate it is a consistent finding that, in paradigms where the cues used at test match those used at study, recognition is superior to recall (see Tulving & Thompson, 1973, for conditions under which recall can be superior to recognition). Without specifying the underlying mechanisms, the key to the difference likely lies in the amount of available information, cues and features between the two test formats. In a recognition setting participant are provided with both the cue and the target item (listed among other distractor items), whereas in a recall setting the participant only has the cue term, making the task harder (for more detail see for example Gillund & Shiffrin, 1984; Haist, Shimamura, & Squire, 1992; Tulving, 1983).

Despite the clear difference between the two memory test formats, participants were not consistently sensitive to what test format they were predicting in their JOL predictions. More specifically, participants changed their responses based on whether they were predicting recognition or recall only when the trial-level design of Experiment 7a was employed and not in the blocked design of Experiment 7b. This was true for both the confidence and binary JOL response formats. In other words, in Experiment 7b participants overall expressed on average the same confidence and gave the same percentage of *yes* JOLs across trials for both recognition and recall predictions. In contrast, in Experiment 7a participants expressed higher average confidence and gave more *yes* JOL predictions on recognition as compared to recall trials, thus aligning the overall predictions more consistently with the subsequent memory performance. This pattern of results shows that delayed JOLs are not truly sensitive to theories about memory as has been demonstrated with immediate JOLs (e.g., Kornell & Bjork, 2009). Rather, it is consistent with the idea that delayed JOLs are relative in nature and might draw on direct comparisons between adjacent trials (as has been already suggested for immediate JOLs, see Koriat, 1997).

The relative nature of metacognitive judgments has been suggested to be characteristic of confidence in particular (Hanczakowski et al., 2013). Besides the immediate JOL, similar ideas have been explored in the context of perceptual metacognitive judgments. Rahnev et al. (2015) demonstrated that a confidence judgment about characteristics of perceptual stimuli (e.g., whether there were more red or blue letters in the display) on any given trial was among other things influenced by confidence on the trial preceding it. Altogether, these ideas link with the suggestion that confidence judgments rank items against each other. In that context it makes sense that on any given trial participants do not only consider the amount of evidence or signal they have access to on that given trial but also consider how it compares to the trial preceding it. The fact that we observed the same pattern of results for confidence and binary responses suggests this relative nature might not be specific to confidence but might be a general characteristic of metacognitive judgments.

That the two response formats were employed in the same way here is not inconsistent with the results of Chapter 4. The results of the previous chapter showed that confidence and binary responses do not necessarily map onto each other directly, consistent with the idea that confidence is not probabilistic (Hanczakowski et al., 2013). However, here we only looked at overall, average use of the confidence scale rather than individual responding. Further, it is likely that anchoring on the confidence scale (i.e. where the likely divide between *yes* and *no* predictions is positioned) can change across tasks (see for example Serra & England, 2012). In general, how the confidence scale is interpreted and used can change within a task if, for example, the make-up of the studied stimuli changes (Zawadzka & Higham, 2016).

As a side note, being a within-subject design, Experiment 7b could also be viewed as comparative in nature. However, the blocks were clearly delineated from each other through instructions slides which were presented before each block, reminding participants what memory test they were making the prediction for and what JOL response format they should employ. As such the design of Experiment 7b did not encourage participants to compare trials within any given block against items in the other blocks. Further, it is likely that participants only consider trials immediately preceding the trial on which the judgment is being made rather than holding in memory all the responses they have given throughout the task. In other words, the fact that no change in responding was observed in Experiment 7b demonstrates that while delayed JOLs are comparative in nature, this is more in terms of comparisons being made in relationship to items immediately preceding the trial at which a judgment is being made rather than in the context of the entire task.

Despite the change in overall responding in Experiment 7a, participants in both experiments overestimated their future recall and underestimated their recognition in terms of calibration. This was true for both response formats. In other words, the change in responding in Experiment 7a did not fully capture the magnitude of difference between recognition and recall performance. This relates to recent findings that confidence reports do not necessarily match onto retrieval probabilities and are relative rather than absolute (Hanczakowski et al., 2013; Zawadzka & Higham, 2015). However, the fact that this was observed with binary responses as well, which are much more clearly separated and defined categories than confidence responses, shows that this is not just a result contingent on the employed response format.

Turning to the signal detection analysis, the results showed that whereas in Experiment 7a participants adopted the same response bias for recognition and recall JOL predictions, in Experiment 7b they adopted a more conservative response bias for recognition as compared to recall predictions. Response bias in the signal detection framework represents the amount of evidence that participants require to give a positive (*yes*) judgment, irrespective of discrimination accuracy. Clearly, the evidence available for a recognition and a recall judgment differs. For example, when presented with the cue, participants might be able to remember the first letter of the target. While this could represent sufficient evidence for a likely subsequent recognition, it is less compelling evidence for successful future recall if no other information is retrieved at time of judgment. In other words, participants are likely to access more evidence for recognition as compared to recall on any given trial and this should lead to more *yes* JOLs for recognition as compared to recall trials. That participants employed the same response bias for recognition and recall JOLs corresponds to the observation that participants gave overall more *yes* and high confidence JOLs on recognition as compared to recall trials. Experiment 7b on the other hand, participants gave the same responses on any given trial irrespective of whether they were predicting recognition or recall even though recognition is more likely. This is why the results show that participants were more conservative in responding on recognition as compared to recall trials. More specifically, in the case of recognition participants gave fewer *yes* predictions (50% of all trials on average) than was warranted given memory performance (80% targets recognised correctly on average) whereas in the case of recall they gave more *yes* predictions (also 50% of all trials on average) than their performance (32% of targets recalled correctly on average) warranted. In other words, these results are consistent with the overall pattern of responding observed in the two experiments. This was seen especially in the binary data but also in the confidence results. In Experiment 7b, we saw a shift if response criterion toward more conservative for recognition predictions on the lower end of the confidence scale.

The only difference between the two response formats was observed with relative JOL accuracy. While the confidence responses for recognition and recall predictions were equally accurate across experiments, the binary responses more accurately predicted recall as compared to recognition in both experiments. That recall predictions might be more accurate is not surprising given the format of the Judgment Phase more closely resembles the format of the memory test for recall than recognition. In both instances only the cue is present. It is likely that in the Judgment Phase participants find it harder to fully account for facilitation in performance by having the target present (among other distractor items).

However, that we would observe this difference on one response format and not the other was not predicted and is hard to account for. While we expected to observe differences in overall responding and calibration across response formats, consistent with other findings in the literature (e.g., Hanczakowski et al., 2013), we had no reason to predict differences in resolution for the two response formats. The two measures employed are different with *d'* relying on parametric assumptions about the underlying distributions of signal not true for AUC. However, the pattern of results observed for *d'* was confirmed when resolution for the binary data was also assessed using AUC. As such this does not seem to explain the difference in the observed results and at the moment we are unable to account for it.

Lastly, we compared recognition and recall JOLs in how accurately they predicted recognition performance. If participants were sensitive to memory test format, it follows that they should more accurately predict recognition performance when that is what they aimed to predict (recognition JOL) as compared to when they were predicting recall performance. In Experiment 7a, participants more accurately predicted future recognition when that was what they were in fact predicting with binary (*yes/no*) responses. This was not the case in Experiment 7b. This is again consistent with the response data suggesting that in Experiment 7b participants did not differentiate between whether they were predicting recognition or recall, rather both JOLs were made in the same way. However, we did not observe any differences in AUC accuracy between experiments and between JOL prediction types, similar to lack of resolution differences in the confidence data discussed above.

Previous chapters have demonstrated, consistent with the literature (e.g., Koriat, 1997) that immediate JOLs are sensitive to mnemonic cues. As described earlier, mnemonic cues are usually stimulus specifics and describe characteristics that generally change across trials. Even without any manipulations, when participants make their JOLs they naturally find some cue words more familiar than others and some targets easier to access than others. The ideas discussed in this chapter similarly highlight that delayed JOLs are likely primarily sensitive to variables that change on a trial level. If metacognitive judgments aim to track relative changes in signal then there must be some variability in the signal encountered. This is why variables that remain consistent across a block of trials are less likely to influence the judgments made. However, results of Chapter 2 showed that in a between subject design it is possible to get differences in JOL responding as participants gave fewer *yes* JOLs when they were presented with pseudo-word cues as compared to real-word cues. Consequently, the most likely candidate for variables that might have an effect on JOL responses across a block of trials are variables that also influence access to the target.

In summary, the current study confirms that delayed JOLs are not sensitive to theories about memory. Clearly only one theory of memory was tested in the context of the present chapter but the results are consistent with findings from studies using other metacognitive paradigms (especially immediate JOLs) that consistently fail to find an effect of explicit theories about memory function on judgments (e.g., Koriat et al., 2004; Kornell & Bjork, 2009). Rather, we have demonstrated that delayed JOLs are comparative and relative in nature. The comparative nature of JOLs has first been discussed in the context of immediate JOLs (Koriat, 1997) and here we show that delayed JOLs might share in this characteristic. It is significant that this was observed across both confidence and binary responses suggesting this might not be specific to confidence judgments but could rather be a general characteristic of metacognitive judgments.

CHAPTER 6

GENERAL DISCUSSION

6.1 Overview

Cognitive processes are accompanied by states of awareness that guide evaluation of their function and content (Fleming et al., 2012; Nelson & Narens, 1990; Overgaard & Sandberg, 2012). This metacognitive ability is understood as an inferential process, evaluating outputs of the cognitive system (Koriat, 2000), that has behavioral consequences (Koriat et al., 2006; Metcalfe & Finn, 2008a). As such, understanding how metacognitive judgments are constructed is crucial. This thesis focused on metacognitive judgments made just after learning and during consolidation, and asked how these judgments relate to the underlying memory processes.

To do this we employed the delayed judgments-of-learning (JOL) paradigm; a prediction of whether recently learned information would be successfully retrieved in the future (Nelson & Dunlosky, 1991). In a typical delayed JOL paradigm participants study cue-target word pairs (Study Phase) following which they are presented with the studied cues and asked to make a prediction (either on a confidence scale; e.g., 0%-20-40-60-80-100% or as a binary *yes/no* judgment) about whether they think they would retrieve the target on the subsequent memory test (Judgment Phase). In contrast to immediate JOLs, delayed JOLs are only made once all items have been studied, rather than on a trial-by-trial basis during study (see Koriat & Bjork, 2006). The last phase of each experiment was a recognition memory test where participants were presented with each of the studied cues and picked the associated target from two or three options (with the exception of Chapter 5 which employed a cued recall test for half of the trials).

The metacognitive literature has employed a range of paradigms to date and one of the key questions throughout has been how the different types of judgments are constructed. This has in particular been the focus of the feeling-of-knowing (FOK) literature (judgments made for temporarily inaccessible items) and the immediate JOL literature (see e.g., Koriat & Levy-Sadot, 2001; Rhodes & Castel, 2008). However, the literature on the delayed JOL has been less rich in this respect (for an exception to this see Metcalfe & Finn, 2008b). The aim of this thesis was to fill this gap by asking how delayed JOLs are constructed.

The focus throughout has been on how delayed JOLs relate to memory. The delayed JOL literature has primarily assumed that the judgment is based on access to information stored in long-term memory (as opposed to short-term memory thought to drive immediate JOLs; for review see Metcalfe & Dunlosky, 2008; Rhodes & Tauber, 2011). In other words, ease of access to the target item was assumed the key determinant – the faster a person can recall the target at time of judgment, the higher the delayed JOL. More recently it has also been shown that at least in the early stages of the judgment process, delayed JOLs can also be influenced by familiarity with the cue used to elicit the judgment (Benjamin, 2005; Metcalfe & Finn, 2008b). The FOK literature has further shown that even access to information that is not strictly the target item under evaluation (such as the learning strategy used to link the target to the cue at study; Hertzog et al., 2014) can also influence the judgment magnitude. This means that, at least in the case of FOK, metacognitive judgments might be a more broad class of evidence evaluation mechanisms where evidence could refer to any seemingly relevant information that comes to mind at time of judgment. This thesis asked whether similar conclusions could be made about the delayed JOL.

Chapters 2 and 3 examined how target- and cue-related manipulations influenced both memory and metamemory. Chapter 4 on the other hand did not manipulate anything and instead asked participants to justify their JOLs. The examination of the content of these responses shed light on how the different memory related information (e.g., cue familiarity and target accessibility) related to different types of JOL predictions. Altogether, these chapters examined the effect of stimulus specific variables (i.e. varied item by item) on delayed JOLs. In contrast, Chapter 5 examined whether JOLs are sensitive to theories about memory i.e. variables consistent across items. This was through asking participants to predict either future recognition or future target recall. Overall, this range of experimental approaches allowed us to address a variety of questions related to the underlying nature of the delayed JOL.

6.2 Summary of findings

Across three experiments, Chapter 2 manipulated the location of the target item on screen and in relation to the cue. Participants were asked to (i) remember target location, (ii) indicate how confident they were that they remembered it accurately and (iii) give a JOL indicating their confidence that they will recognise the target on a subsequent memory test. The results indicated that participants' confidence that they remembered target location was directly related to their JOL confidence that they will also recognise it. Further, it was shown that the accuracy of memory for target location also influenced JOLs (with higher JOLs if target location was remembered accurately) but only when memory for target location was also related to memory for the target (i.e. if it was recognised). Firstly, this shows that delayed JOLs are related to the quantity of information that is accessible at time of judgment and that this information does not need to be specifically about what the target item is (i.e. semantic and orthographic information). This relates the delayed JOL to recent similar findings in the

FOK literature (Brewer et al., 2010; Hertzog et al., 2014). A question that arises from these findings is whether metacognitive judgments are not just sensitive to the amount of information that is accessible at time of judgment but whether they might also be sensitive to the amount of information that is *potentially* accessible. If it is indeed quantity of information that matters the most in metacognitive judgments then it is possible that judgments for items, which have more information associated with them (even if this additional information is not relevant to what is being predicted), might receive more confident judgments than items which are not accompanied by additional information. This has already been shown to a certain degree in TOT experiences and FOKs where target items presented alongside pictures at study (as compared to items presented alone or only accompanied by other words) received more TOT reports and higher FOKs (Schwartz et al., 2014; Schwartz & Smith, 1997).

Further, we found that it is not just quantity but also quality of the accessed information (i.e. how accurately it has been remembered) that relates to the magnitude of a delayed JOL. This is contrary to suggestions in the literature that accuracy of the accessed information does not play a role in metacognitive judgments (Koriat, 2000). Thomas et al. (2012) similarly observed that accuracy of access to semantic but not perceptual (font colour) features of the target influenced FOKs. Based on these results, they suggested that the accuracy of the accessed target-related partial features only influences judgments when it is semantic in nature and not when it is perceptual. However, the results presented here rather suggest that the determining factor in whether the accuracy of the accessed partial features is related to metamemory judgments is whether this access is in turn related to target memory. In other words, when accurate access to the target-related feature (here spatial position of the target) was related to accuracy of recognition memory for the target, then it impacted JOL magnitude. It is likely that these kinds of effects are most indicative of the quality of encoding of the target item and the binding between its constituent features. One would expect to observe these effects for target-related information that is closely bound to the target item i.e. in instances where the memory for that information also relates to memory for target identity as that should lead to a richer retrieval experience. When participants can clearly remember the target item (including simultaneous accessing what and where it was), they are more likely to give a high JOL.

Chapter 3, in contrast, looked at the role of the cue in delayed JOLs. More specifically we investigated two forms of cue familiarity and how they impacted memory and metamemory. Firstly, we manipulated pre-experimental cue familiarity across the two experiments with Experiment 4a employing pseudo-word cues (unknown to participants) and Experiment 4b employing real-word cues (familiar words that participants would have encountered in their real-life). Further, both experiments manipulated experimental cue-familiarity by exposing participants to half the cues in a rating task preceding the Study Phase. We investigated how both types of familiarity impacted memory and metamemory but this time using binary (*yes/no*) JOLs.

Firstly, we found that both types of familiarity improved target memory and that these effects interacted. More specifically, memory for targets paired with real-word (i.e. pre-experimentally familiar) cues was better than memory for targets paired with pseudo-word cues. However, experimentally manipulating familiarity of the pseudo-word cues also improved memory. Memory for targets paired with experimentally familiar pseudo-word cues was the same as memory for targets paired with real-word cues. As such experimentally manipulating familiarity can provide an encoding advantage similar to that offered by pre-experimental familiarity with the to-be-studied material.

Further, we found that both types of familiarity influenced JOLs and that these effects also interacted. The metamemory results showed that both types of familiarity influenced JOLs and, contrary to the memory results, did so in an additive manner. The effect of experimental familiarity on JOLs was bigger with pseudo-word cues as compared to real-word cues, but both effects were significant. In other words, even though experimental familiarity did not have a significant effect on memory performance in the case of real-word cues, it still influenced JOL responding. It is clear that while variables that influence memory also influence metamemory, metamemory judgments are also sensitive to influences that do not necessarily correspond to changes in memory performance.

Both Chapter 2 and 3 can be seen as an extension of the classic literature on cue- and target-related effects in the metamemory literature (e.g., Koriat & Levy-Sadot, 2001; Metcalfe & Finn, 2008b). Chapter 4 also adds to this literature but from a different angle. Instead of manipulating the nature of the cue or the target, participants completed a standard delayed JOL task without any variables being manipulated. On a subset of the JOL trials, participants wrote justifications for the response they have given. They were not given any instructions on how to write the justifications. The first finding of this experiment was that participants referenced both cue familiarity and target accessibility in their justifications. This therefore supports other findings in the literature and shows that these effects are not simply a result of making certain characteristics of the cue or the target more salient to participants through varying their strength across experimental trials. The real strength of this approach is that nothing was manipulated and no instructions were given to participants – in that sense their justifications were entirely spontaneous.

Further, in Experiment 5 of Chapter 4, participants gave JOLs on a confidence scale (0%-20%-40%-60%-80%-100%) whereas in Experiment 6 of Chapter 4 they gave binary (*yes/no*) JOLs followed by three-point confidence about that judgment (*sure-maybe-guess*). This means that in both experiments participants could make six distinct JOLs with accompanying confidence. However, only in Experiment 6 was there a clear *yes/no* distinction between the 6

JOLs available to participants. The intuitive assumption is that the confidence scale can be split down the middle into an equal number of *yes* and *no* responses (see e.g., Mason & Rottello, 2009). For example 40% confidence suggests a higher likelihood of failure than success in target recognition and so probabilistically could be interpreted as a *no* prediction. Further, some have argued that there is an underlying *yes/no* distinction in confidence judgments with participants first making the binary judgment before assigning confidence (e.g., Dunlosky et al., 2005) and that this division should be along the centre of the confidence scale intuitively makes sense (see e.g., Hanczakowski et al., 2013). However, the pattern of justifications between the two response formats differed and there was not a clear one-to-one mapping in the justifications for the two response formats. More specifically, it seems that the difference in justifications between the 0% and 20% confidence most closely resembled the differences between *yes* and *no* justifications. Further, whereas the different levels of *yes* justifications were fairly well defined, there were more similarities than differences across the different levels of *no* JOL predictions.

At the very least this data shows that we cannot always make assumptions about how a specific response format is interpreted beyond what we know has been offered to participants. In other words, if participants are given a confidence percentage scale, it is possible that there is an underlying *yes/no* judgment in terms of which the scale can be interpreted but we need to confirm this, we cannot just assume it. Even more importantly, we cannot assume where such a *yes/no* distinction might be located on the confidence scale. This is consistent with findings that confidence judgments are unlikely to be probabilistic (Hanczakowski et al., 2013) and that the use of the confidence scale is flexible. More specifically, participants' use of the confidence scale changes if they are asked whether they will 'remember' or whether they will 'forget' the target word (Serra & England, 2012) and even one participant can, within one experimental session, adjust their use of the confidence scale and the distance between the separate points on the scale if the make-up of the items changes, such as through adding a new set of hard or easy items to those already studied (Zawadzka & Higham, 2016).

Lastly, we saw that participants' JOLs were not sensitive to whether they were predicting recognition or recall. Both experiments in Chapter 5 (Experiments 7a and 7b) employed a within-subject design comparing recognition and recall JOL predictions but whereas in Experiment 7a the predictions were interspersed on a trial-level, in Experiment 7b they were blocked. Consistent with the literature (e.g., MacDougall, 1904) and our expectations, recognition performance was significantly better than recall performance. For participants to show sensitivity to memory test predicted, they would need to give fewer *yes* and lower confidence JOLs for recall as compared to recognition predictions. The results showed that participants changed their JOLs (both JOL confidence magnitude and percentage of *yes* responses in binary JOLs) with the memory test they were predicting only in the trial-level (Experiment 7a) and not the blocked (Experiment 7b) design. This finding has two major implications. Firstly, it seems likely that delayed JOLs are *not* sensitive to theory-based processes (as compared to stimulus specific influences as seen in the previous chapters). Secondly, it suggests that delayed JOLs are a relative rather than an absolute judgment. In other words, participants might be making a judgment on any one trial in comparison with the trials immediately preceding it (it is unlikely that any given trial is compared to *all* other trials consistent with the results of Experiment 7b). In other words, the amount of evidence accessible on any given trial is likely compared to the evidence accessible on at least the preceding trial – if it is higher then the participant is likely to also give higher confidence JOL. In the context of the present experiments, if the memory test for which predictions are made changes on a trial level (Experiment 7a), such that participants might predict recognition on the first trial and recall on the subsequent trial, then participants will incorporate that into the value of the final judgment outputted. It is noteworthy that this was observed both with confidence and binary JOLs. This is important because while confidence has already been proposed to be relative rather than absolute (e.g., Hanczakowski et al., 2013) a similar argument has not been made about binary judgments in the metacognitive literature.

This idea is consistent with recent findings relating to other metacognitive paradigms but to date has not been explored with delayed JOLs. For example, immediate JOLs have been found to be sensitive to a number of influences (e.g., fluency of processing) only in within-subject as compared to between subject designs (e.g., Susser et al., 2013). This is further related to the finding that confidence (at least in immediate JOL tasks) is relative rather than absolute with participants meaningfully ranking the items against each other (in a comparative way) rather than attempting to express exact probability of future retrieval success (e.g., Hanczakowski et al., 2013). Lastly, a study from the perceptual metacognitive literature has found that confidence judgments on any one trial were also influenced (i) by unrelated confidence judgments on the same trial and (ii) by confidence expressed on the preceding trial (Rahnev et al., 2015). The important finding of this study was that the observed relationship between judgments were not just results of differing strength of the stimulus and attention, factors which were controlled for across experiments and in the regression analyses employed. Rather, the researchers concluded that, at least in metacognitive judgments about perception, participants evaluate the stimulus related signal in the context of the strength of signal on the previous trials. These confidence leaks are yet another example of confidence being relative rather than absolute with the researchers stressing that judgments on any given trial are not made in isolation from the context of the task and the experiences of the preceding trials. It is possible that the relative nature of judgments might be true for metamemory, and indeed metacognitive judgments in general. Even cognitive judgments such as those made in recognition memory tasks (i.e. is an item *old* or *new*) were shown to be relative with responses on any one trial also influenced by responses on preceding trials (Malmberg & Annis, 2012).

In summary, the delayed JOL can be seen as a general evidence evaluation process. It is the quantity of the accessed information at time of judgment that determines the JOL value; information here refers to any seemingly relevant accessible information at time of judgment. Further, the evidence quality (i.e. its accuracy) can also influence judgments. Participants are capable of justifying their judgments and do so with reference to the underlying memory processes. JOLs are determined primarily by stimulus specific variables that vary on a trial-level that together constitute a signal which participants evaluate as either indicative of future target retrieval or not. Given the JOL is based on evaluating changes in signal, between subject variables are unlikely to influence the judgment unless they also influence access to the target. For example, the pre-experimental familiarity manipulation in Chapter 3 influenced JOLs outputted between experiments because participants in the pseudo-word cue condition found it harder to remember the targets which led to overall fewer *yes* JOLs. Otherwise it is primarily a comparative judgment, and so will be influenced not just by the quantity and quality of the evidence accessible at time of judgment but also how that evidence compares to what was available on preceding trials. It is even possible participants might recalibrate their responses throughout the task as the available evidence changes.

6.3 Methodological implications

Both Chapter 4 and Chapter 5 used confidence and binary JOLs, allowing for direct comparisons. In Chapter 4 we collected justifications for both types of JOL responses whereas in Chapter 5 we investigated whether they give the same pattern of results for the experimental manipulations under investigation (namely whether participants predicted future recognition or recall and whether this was manipulated on a trial-level or in a blocked design). Chapter 5 did not observe any differences between confidence and binary JOLs in that they gave overall the same pattern of results as far as the key main effects of interest were concerned. Nevertheless, there was a three-way interaction between all factors of interest indicating that there were slight variations in how the two response formats were employed and in some instances the proportion of *yes* responses was higher than the average JOL confidence expressed. This is consistent with Chapter 4 findings that, if there is a *yes/no* binary split in the confidence scale, it is most likely located lower on the scale than its mid-point. Chapter 4 overall indicated that there is not necessarily a 1-to-1 mapping between the two response formats. It is likely that how participants interpret confidence changes between experiments and even participants. It is possible that in some situations participants split the confidence scale down the middle into *yes* and *no* responses. But the data presented in this thesis demonstrate that this interpretation cannot be taken for granted. While the results of Chapter 4 are consistent with the notion that there is some underlying split into *yes* and *no* judgments in the confidence scale (even if most of the confidence scale seems to correspond to *yes* predictions and only the lowest responses could be interpreted as *no* predictions), that assumption also needs to be confirmed with further data.

The take away message is that researchers need to consider more closely the assumptions underlying the paradigm and response format they employ. Binary and confidence judgments should not be considered interchangeable. Further, confidence should not be interpreted with further meaning than that which is provided to participants unless participants explicitly indicate or are instructed to interpret it in that manner (e.g., interpreting a subset of responses as corresponding to a *yes* prediction). Further, results should be confirmed across multiple response formats (as suggested by Hanczakowski et al., 2013). In a series of experiments independent of this thesis we have found that participants' likelihood of reporting déjà vu or TOT experiences in an experimental setting is influenced by how they are asked about the experience, irrespective of the experimental paradigm adopted (Jersakova et al., 2016). This means some studies could falsely conclude they have recreated these subjective experiences in the laboratory when in fact the pattern of results they have obtained is rather a direct consequence of the format of responding they have employed. Researchers need to ensure a pattern of results is specific to the psychological phenomenon under investigation rather than emerging from the response format employed.

As Zawadzka & Higham (2015) point out, each response format has its benefits. Confidence judgments provide more of a range of responses and a corresponding increase in sensitivity. It is clear that metacognitive experiences are rarely all or nothing and it is often the medium confidence responses (outside of 0% and 100% confidence) that are the most interesting from a research perspective. Binary responses on the other hand are easier to interpret, as there is less likelihood that participants might use them in ways that is not in line with how the researchers assume they are being used.

Lastly, Chapter 4 introduced a novel approach to the study of metacognition. While the idea of asking participants to give written reports in metacognitive tasks is not entirely novel (e.g., Koriat, 1980), it has not yet been used in this manner where participants are asked to justify their trial-level responses. Further, we combined this approach with natural language processing techniques and machine learning analysis of these reports is, which allowed for a quantitative approach to otherwise qualitative data. Firstly, this study adds to the few studies that have shown that asking participants to report on their strategies and approach to the task can be highly informative. Already Eagle (1967) noted that participants' performance on a memory task was more closely related to their reported learning strategies as compared to the strategies they were instructed to use by the experimenter. Altogether this helps to highlight that asking participants for detailed self-reports can elucidate processes that might otherwise be overlooked.

6.4 Future directions

Much of the metamemory literature has separately investigated the extent to which a given judgment is sensitive to the manipulation of a specific variable (e.g., cue familiarity). Consequently, there is now a growing body of evidence indicating that the type of accessed partial information that can impact metamemory predictions is truly varied, ranging from semantic and perceptual features of the target (Koriat et al., 2003; Thomas et al., 2012) to other elements present at time of encoding (Schwartz et al., 2014) or aspects of encoding strategies (Hertzog et al., 2014) as well as spatial information as shown in Chapter 2. An obvious next step is to bring these variables together in a comprehensive study and investigate whether certain types of information have more of an influence on metamemory judgments as compared to others and whether certain influences still have a significant impact of metacognition once other variables are taken into account. This should be done in the context of first identifying *types* or *categories* of influences and manipulations and then directly evaluating them against each other. A related but separate point is that we should also move toward more complex material than cue-target word pairs.

The consideration of the level to which metacognitive judgments are relative calls for investigation not just as a way to find possible links across metacognitive judgments but also on its own merits due to its methodological and theoretical implications. Researchers are used to considering responses in isolation and yet the literature on confidence leaks (e.g., Rahnev et al., 2015) and some of the results presented here suggest this might not be correct. Linked to this is the more general question of how the context in which the judgment is being made influences the outcome of that judgment. This is related to findings that metacognitive judgments are sensitive to question framing (e.g., Finn, 2008). Overall, this comes back to the idea that there are assumptions in the way metacognitive and cognitive tasks are applied and interpreted in the literature that might not be valid. This is a methodological but also a theoretical point. Understanding better how the experimental set-up influences responding leads to a clearer understanding of the data collected. It also improves our understanding of metacognition and the extent to which judgments are context sensitive. Mapping the variables related to other stimuli encountered and responses given in an experimental setting allows for a fuller understanding of how metacognitive judgments are made.

Throughout this thesis we have borrowed from some of the literature on FOK and immediate JOLs in developing avenues of research for better understanding the basis of delayed JOLs. The obvious next step and something still not addressed sufficiently in the literature is the direct comparison between these paradigms. Most attention has been given to comparing the immediate and delayed JOL paradigms, mainly with the focus on understanding why the latter is more accurate than the former (see e.g., Koriat, 1997; Metcalfe & Dunlosky, 2008; Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011). This has led to a range of theories, the core of which have primarily focused on the idea that in the delayed JOL participants have access to cues that are more diagnostic of future memory performance than is the case in the immediate JOL. Some work has also been done on comparing the delayed JOL and FOK in older adults, suggesting they are not equivalent (Souchay & Insingrini, 2012). Other than those efforts, the paradigms have not truly been systematically compared. Clearly each judgment relates to a different aspect of the memory process (encoding, consolidation and retrieval) and is made under different circumstances, which means there are clear differences between them. What is interesting however is to ask what unities them, whether there is such a thing as an underlying similarity that could be considered the core of metamemory judgments.

Related to the above, the next question and something entirely missing from the field is a comparison of metamemory paradigms to other metacognitive phenomena. Again, it is clear that each metacognitive judgment relates closely to the underlying cognitive process that it is monitoring and along these lines there will be clear differences. But it is of interest to explore in more detail whether there are underlying similarities as the use of an overarching term ‘metacognition’ would suggest. For example, metacognition (especially in terms of memory judgments) has been related to executive functions and frontal lobe processes (see Shimamura, 2000) so that is one possible avenue of research. That said, some research has found that memory performance was a better predictor of metamemory accuracy than executive function scores in Azheimer’s patients (Souchay et al., 2002). This would suggest that the link to executive functions could remain a minor one but this needs to be investigated in more detail, especially with reference to other types of metacognitive judgments.

An alternative to a neuropsychological approach would be a modelling approach. If all metacognitive judgments can be generalised as evidence evaluation processes where evidence refers to the strength of the signal provided by the cognitive system under consideration then it should be possible to come up with a generalized computational model of how that evidence evaluation feeds into the metacognitive judgment. It is clear that the evidence available in judgments on a memory as compared to a perceptual task will vary widely. Of interest here would be to explore whether there are characteristics of the evaluation processes, such as their relative nature for example, that would generalise across metacognitive domains. It is possible that a general class of influences (e.g., stimulus specific, context specific etc.) that could generalise across metacognitive tasks could be identified. Once general classes of influences are identified, a general modelling approach (e.g., drawing on artificial neural networks) becomes feasible.

6.5 Conclusions

Delayed JOLs have been traditionally treated as a special case of the immediate JOL. Consequently, theories outlining the underlying mechanisms of delayed JOLs have focused on how the paradigm differs to immediate JOLs, such as positing that the delayed JOL is better positioned to assess the strength of access to the target (for reviews see Metcalfe & Dunlosky, 2008; Rhodes & Tauber, 2011). Correspondingly, the delayed JOL literature has treated the paradigm as an absolute judgment, each trial individually assessing the level of access to the target item and based on that evaluation computing the probability of subsequent target retrieval on the upcoming memory test.

That target accessibility is a determinant in delayed JOLs is clearly established in the literature and the results reported here support that. However, across seven experiments, this thesis demonstrated that delayed JOLs are also subject to a *range* of influences, beyond the level of memory for target identity (i.e. *what* the target is). Variables that impact memory performance as indexed by target recognition (e.g., pre-experimental cue familiarity) also impact JOLs. However, in some instances variables which do no impact memory performance still influence JOLs. We saw this for example when participants' confidence that they remembered target location significantly influenced their JOLs, even in instances when access to this information did not relate to memory for the target (Experiment 3, Chapter 2).

When groups differ in memory performance, this can lead to absolute changes in JOL responding between the two groups (at least in binary JOLs, see Chapter 3). Nevertheless, another novel contribution of this thesis is the demonstration that the delayed JOL is primarily a relative judgment sensitive to variables that vary on a trial-level basis (see Chapter 5). Rather than being an absolute assessment of retrieval probability, the delayed

JOL can be characterised an evaluation of signal strength in relation to other experimental trials and to the experimental context. It is possible that target accessibility accounts for primary, substantial, coarse changes in responding while the trial-level, relative fluctuations might lead to judgment fine-tuning.

Another contribution of this thesis is methodological. The interest in metacognition stems from its vital role in elucidating the nature of healthy cognitive development as well as improving our understanding of cognitive impairment. To do so, we need to have a good grasp of the strengths and limitations of the paradigms we employ. This is especially true when it comes to interpretation of results. That a judgment might change with what it is predicting (e.g., recognition vs. recall) when manipulated on a trial-level does not necessarily imply that that judgment is sensitive to theory-based processes (Chapter 5). Similarly, that a participant gives a below 50% level of confidence on a given trial does not necessarily mean that they are predicting they will not retrieve the target (Chapter 4). Rather it means participants are indicating uncertainty, which is vastly different. Such, often implicit, assumptions in how data is collected and analysed can lead to incorrect conclusions about the processes under investigation. As such it is important to understand the underlying assumptions of the paradigms we employ and to check for their validity.

Future metacognitive research needs to investigate fundamental methodological questions in more detail. Further, the metacognitive literature needs to start asking bigger, overarching questions. This should also be linked to a concentrated replication effort of key findings in the literature. As in so many fields in psychology today, it is time to start identifying core themes and theories and ultimately to try to develop a comprehensive account of metacognition.

REFERENCES

Ackerman, R., & Thompson, V. (2014). Meta-reasoning: What can we learn from meta-memory? In A. Feeney & V. Thompson (Eds.), *Reasoning as Memory*. Hove, UK: Psychology Press.

Alban, M. W., & Kelley, C. M. (2013). Embodiment meets metamemory: weight as a cue for metacognitive judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1628–34. doi: 10.1037/a0032420

Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, *55*, 64–88. doi:10.1016/j.jml.2006.02.002

Allen, R. J. (2015). Memory binding. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioural Sciences* (2nd ed., pp. 140–146). Elsevier Sciente Ltd.

Allen, R. J., Havelka, J., Falcon, T., Evans, S., & Darling, S. (2015). Modality specificity and integration in working memory: Insights from visuospatial bootstrapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 820–830. doi:10.1037/xlm0000058

Allen, R. J., Vargha-Khadem, F., & Baddeley, A. D. (2014). Item-location binding in working memory: Is it hippocampus-dependent? *Neuropsychologia*, *59*, 74–84. doi:10.1016/j.neuropsychologia.2014.04.013

Arango-Muñoz, S. (2010). Two levels of metacognition. *Philosophia*, *39*(1), 71–82. doi:10.1007/s11406-010-9279-0

Arango-Muñoz, S. (2013). The nature of epistemic feelings. *Philosophical Psychology*, *27*(2), 193–211. doi:10.1080/09515089.2012.732002

Badham, S. P., & Maylor, E. A. (2011). Age-related associative deficits are absent with nonwords. *Psychology and Aging*, *26*(3), 689–694. doi:10.1037/a0022205

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/BF03193014

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bein, O., Livneh, N., Reggev, N., Gilead, M., Goshen-Gottstein, Y., & Maril, A. (2015). Delineating the effect of semantic congruency on episodic memory: The role of integration and relatedness. *PLoS ONE*, *10*(2), e0115624. doi:10.1371/journal.pone.0115624

Benjamin, A. S. (2005). Response speeding mediates the contributions of cue familiarity and target retrievability to metamnemonic judgments. *Psychonomic Bulletin & Review*, *12*(5), 874–879. doi:10.3758/BF03196779

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, *9*, 529–533. doi:10.1016/S0022-5371(70)80096-2

Bower, G. H., & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, *20*(2), 119–120. doi:10.3758/BF03335632

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 717–726. doi:10.1016/S0022-5371(72)80006-9

Brewer, G. A., Marsh, R. L., Clark-Foos, A., & Meeks, J. T. (2010). Noncriterial recollection influences metacognitive monitoring and control processes. *Quarterly Journal of Experimental Psychology (2006)*, *63*(10), 1936–42. doi:10.1080/17470210903551638

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parker, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, 1–16. doi:10.1111/j.2044-8260.1982.tb01421.x

Brown, A. S. (2004). *The Déjà Vu Experience* (1st ed.). Hove: Psychology Press.

Brown, A. S. (2012). *The Tip of the Tongue State* (1st ed.). Hove: Psychology Press.

Burgess, N., Becker, S., King, J. A., & Keefe, J. O. (2001). Memory for events and their spatial context : models and experiments. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *356*, 1493–1503. doi:10.1098/rstb.2001.0948

Burgess, N., Maguire, E. A., & Keefe, J. O. (2002). The Human hippocampus and spatial and episodic memory. *Neuron*, *35*, 625–641. doi:10.1016/S0896-6273(02)00830-9

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review*, *14*(1), 107–11. doi:10.3758/BF03194036

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81. doi:10.1016/0010-0285(73)90004-2

Chua, E. F., Hannula, D. E., & Ranganath, C. (2012). Distinguishing highly confident accurate and inaccurate memory: Insights about relevant and irrelevant influences on memory confidence. *Memory*, *20*(1), 48–62. doi:10.1080/09658211.2011.633919

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing : A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684. doi:10.1016/S0022-5371(72)80001-X

Darling, S., Allen, R. J., Havelka, J., Campbell, A., & Rattray, E. (2012). Visuospatial bootstrapping: Long-term memory representations are necessary for implicit binding of verbal and visuospatial working memory. *Psychonomic Bulletin & Review*, *19*(2), 258–63. doi:10.3758/s13423-011-0197-3

de Bruin, A. B. H., Rikers, R., & Schmidt, H. G. (2007). Improving metacomprehension accuracy and self-regulation in cognitive skill acquisition: The effect of learner expertise. *European Journal of Cognitive Psychology*, *19*(4–5), 671–688. doi:10.1080/09541440701326204

de Sousa, R. (2009). Epistemic feelings. *Mind and Matter*, *7*(2), 139–161.

Dennis, S. (2006). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57–70). London: Routledge.

Descartes, R. (1911). Meditations on first philosophy (E. S. Haldane Trans.). In *The Philosophical Works of Descartes*. Cambridge: Cambridge University Press (original work published 1641).

DeWitt, M. R., Knight, J. B., Hicks, J. L., & Ball, B. H. (2012). The effects of prior knowledge on the encoding of episodic contextual details. *Psychonomic Bulletin & Review*, *19*(2), 251–257. doi:10.3758/s13423-011-0196-4

Dixon, R. A., Hultsch, D. F., & Hertzog, C. (1988). The Metamemory in Adulthood (MIA) questionnaire. *Psychopharmacological Bulletin*, *24*(4), 671–688.

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language*, *36*, 34–49. doi:10.1006/jmla.1996.2476

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, *52*, 551–565. doi:10.1016/j.jml.2005.01.011

Eagle, M., & Leiter, E. (1964). Recall and recognition in intentional and incidental learning. *Journal of Experimental Psychology*, *68*(1), 58–63. doi:10.1037/h0044655

Eagle, M. N. (1967). The effect of learning strategies upon free recall. *American Journal of Psychology*, *80*(3), 421–425.

Ebbinghaus, H. (1964). *Memory: A Contribution to Experimental Psychology* (H. A. Ruger, C. E. Bussenius & E. R. Hilgard. New York, NY: Dover Publications (Original work published 1885).

Ernst, A., Ernst, A., Moulin, C. J. A., Souchay, C., Mograbi, D. C., & Morris, R. (2016). Anosognosia and metacognition in Alzheimer's Disease. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. Oxford, UK: Oxford University Press. doi:10.1093/oxfordhb/9780199336746.013.12

Eysenck, M. W. (1979). The feeling of knowing a word's meaning. *British Journal of Psychology*, *70*, 243–251. doi:10.1111/j.2044-8295.1979.tb01681.x

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition*, *36*(4), 813–821. doi:10.3758/MC.36.4.813

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *33*(1), 238–244. doi:10.1037/0278-7393.33.1.238

Flavell, J. H. (1971). First discussant's comments: What is memory development in the development of? *Human Development*, *14*, 272–278. doi:10.1159/000271221

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition*. Hillsdale, NJ: Erlbaum.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–6. doi:10.1098/rstb.2012.0021

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. *Psychological Science*, *26*(1), 89–98. doi:10.1177/0956797614557697

Fletcher, L., & Carruthers, P. (2012). Metacognition and reasoning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1366–1378. doi:10.1098/rstb.2011.0413

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, *1*(3), 213–216. doi:10.3758/BF03198098

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition*, *7*(7), 1–26. doi:10.1006/ccog.1997.0321

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models : A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. doi:10.1037//0033-295X.98.4.506

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67. doi:10.1037/0033-295X.91.1.1

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5). doi:10.1037/0278-7393.7.5.311

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732–764. doi:10.1007/978-1-4612-9995-0_1

Haist, F., Shimamura, A., & Squire, L. (1992). On the relationship between recall and recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *18*(4), 691–702. doi: 10.1037/0278-7393.18.4.691

Hamel, L. H. (2009). *Knowledge discovery with support vector machine*. Hoboken, NJ: Wiley.

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. a. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language*, *69*(3), 429–444. doi:10.1016/j.jml.2013.05.003

Hart, J. T. (1965). Memory and the feeling of knowing experience. *Journal of Educational Psychology*, *56*(4), 208–216. doi:10.1037/h0022263

Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, *20*, 1315–1326. doi:10.1002/hipo.20857

Herfurth, K., Kasper, B., Schwarz, M., Stefan, H., & Pauli, E. (2010). Autobiographical memory in temporal lobe epilepsy: role of hippocampal and temporal lateral structures. *Epilepsy & Behavior : E&B*, *19*(3), 365–71. doi:10.1016/j.yebeh.2010.07.012

Hertzog, C., Fulton, E. K., Sinclair, S. M., & Dunlosky, J. (2014). Recalled aspects of original encoding strategies influence episodic feelings of knowing. *Memory & Cognition*, *42*, 126–40. doi:10.3758/s13421-013-0348-z

Higham, P. A. (2011). Accuracy discrimination and the type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of Remembering and Metacognition*. New York, NY: Palgrave Macmillan.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. doi:10.1037/0033-295X.95.4.528

Horner, A. J., & Burgess, N. (2013). The associative structure of memory for multi-element events. *Journal of Experimental Psychology. General*, *142*(4), 1370–83. doi:10.1037/a0033626

Horner, A. J., & Burgess, N. (2014). Pattern completion in multielement event engrams.

Current Biology, *24*(9), 988–92. doi:10.1016/j.cub.2014.03.012

Howard, C. E., Andrés, P., Broks, P., Noad, R., Sadler, M., Coker, D., & Mazzoni, G. (2010). Memory, metamemory and their dissociation in temporal lobe epilepsy. *Neuropsychologia*, *48*(4), 921–32. doi:10.1016/j.neuropsychologia.2009.11.011

Illman, N. A., Kemp, S., Souchay, C., Morris, R. G., & Moulin, C. J. A. (2016). Assessing a metacognitive account of associative memory impairments in temporal lobe epilepsy. *Epilepsy Research and Treatment*, *2016*, Article ID 6746938, 11 pages. doi:10.1155/2016/6746938

James, W. (1890). *Principles of psychology*. New York, NY: Holt.

Jameson, K. A., Narens, L., Goldfarb, K., & Nelson, T. (1990). The influence of near-threshold priming on metamemory and recall. *Acta Psychologica*, *73*, 55–68. doi:10.1016/0001-6918(90)90058-N

Jersakova, R., Moulin, C. J. A., & O'Connor, A. R. (2016). Investigating the role of assessment method on reports of déjà vu and tip-of-the-tongue states during standard recognition tests. *PloS ONE*, *11*(4), e0154334. doi:10.1371/journal.pone.0154334

Jersakova, R., Souchay, C., & Allen, R. J. (2015). Negative affect does not impact semantic retrieval failure monitoring. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Expérimentale*, *69*(4), 314–326. doi:10.1037/cep0000065

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language*, *48*(4), 704–721. doi:10.1016/S0749-596X(02)00504-1

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1322–1337. doi:10.1098/rstb.2012.0037

Köhler, S., Moscovitch, M., & Melo, B. (2001). Episodic memory for object location versus episodic memory for object identity : Do they rely on distinct encoding processes ? *Memory & Cognition*, *29*(7), 948–959. doi:10.3758/BF03195757

Konkel, A., & Cohen, N. J. (2009). Relational memory and the hippocampus: representations and methods. *Frontiers in Neuroscience*, *3*(2), 166–74. doi:10.3389/neuro.01.023.2009

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639. doi:10.1037/0033-295X.100.4.609

Koriat, A. (1997). Monitoring one’s own knowledge during study : A cue-utilization approach to judgments of learning, *126*(4), 349–370. doi:10.1037/0096-3445.126.4.349

Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*, 149–171. doi:10.1006/ccog.2000.0433

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(2), 187–94. doi:10.1037/0278-7393.31.2.187

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners’ sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–72. doi:10.3758/BF03193244

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one’s own forgetting: the role of experience-based and theory-based processes. *Journal of Experimental Psychology. General*, *133*(4), 643–56. doi:10.1037/0096-3445.133.4.643

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of cue familiarity and accessibility heuristics to feeling of knowing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*(1), 34–53. doi:10.1037//0278-7393.27.1.34

Koriat, A., Levy-Sadot, R., Edry, E., & de Marcas, S. (2003). What do we know about what we cannot remember? Accessing the semantic attributes of words that cannot be recalled. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1095–105. doi:10.1037/0278-7393.29.6.1095

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for Confidence. *Journal of Experimental Psychology : Human Learning and Memory*, *6*(2), 107–118. doi:10.1037/0278-7393.6.2.107

Koriat, A., & Lieblich, I. (1977). A study of memory pointers. *Acta Psychologica*, *41*, 151–164. doi:10.1016/0001-6918(77)90032-4

Koriat, A., Ma’ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition : Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36–69. doi:10.1037/0096-3445.135.1.36

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*(3), 595–608. doi:10.1037/0278-7393.32.3.595

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves : Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*(2), 147–162. doi:10.1037//0096-3445.131.2.147

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. doi:10.1037/a0017350

Leritz, E. C., Grande, L. J., & Bauer, R. M. (2006). Temporal lobe epilepsy as a model to understand human memory: The distinction between explicit and implicit memory. *Epilepsy & Behavior*, *9*(1), 1–13. doi:10.1016/j.yebeh.2006.04.012

Liu, Y., Su, Y., Xu, G., & Chan, R. C. K. (2007). Two dissociable aspects of feeling-of-knowing: knowing that you know and knowing that you do not know. *Quarterly Journal of Experimental Psychology (2006)*, *60*(5), 672–80. doi:10.1080/17470210601184039

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*(1), 149–157. doi:10.1037/0278-7393.16.1.149

MacDougall, R. (1904). Recognition and recall. *Journal of Philosophy*, *1*(9), 229–233. doi:10.2307/2010991

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide*. Hove, UK: Psychology Press.

Maguire, E. A., & Mullally, S. L. (2013). The hippocampus : A manifesto for change. *Journal of Experimental Psychology: General*, *142*(4), 1180–1189.

doi:10.1037/a0033650

Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General*, *141*(2), 233–259. doi:10.1037/a0025277

Marquié, J.-C., & Huet, N. (2000). Age differences in feeling-of-knowing and confidence judgments as a function of knowledge domain. *Psychology and Aging*, *15*(3), 451–460. doi:10.1037/0882-7974.15.3.451

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 509–27. doi:10.1037/a0014876

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General*, *122*(1), 47–60. doi:10.1037/0096-3445.122.1.47

McGlynn, S. M., & Schacter, D. L. (1989). Unawareness of deficits in neuropsychological syndromes. *Journal of Clinical and Experimental Neuropsychology*, *11*(2), 143–205. doi:10.1080/01688638908400882

Metcalfe, J., & Dunlosky, J. (2008). Metamemory. In H. L. Roediger (Ed.), *Learning and Memory: A Comprehensive Reference* (pp. 349–362). Oxford, UK: Elsevier.

Metcalfe, J., Eich, T. S., & Miele, D. B. (2013). Metacognition of agency : proximal action and distal outcome. *Experimental Brain Research*, *229*, 485–496. doi:10.1007/s00221-012-3371-6

Metcalfe, J., & Finn, B. (2008a). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. doi:10.3758/PBR.15.1.174

Metcalfe, J., & Finn, B. (2008b). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(5), 1084–97. doi:10.1037/a0012580

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(4), 851–61. doi:10.1037//0278-7393.19.4.851

Metcalfe, J., & Son, L. (2012). Anoetic, noetic and autonoetic metacognition. In M. Beran, J. R. Brandl, J. Perner, & J. Proust (Eds.), *The Foundations of Metacognition*. Oxford, UK: Oxford University Press.

Mill, R. D., & O'Connor, A. R. (2014). Question format shifts bias away from the emphasised response in tests of recognition memory. *Consciousness and Cognition*, *30*, 91–104. doi:10.1016/j.concog.2014.09.006

Morson, S. M., Moulin, C. J. A., & Souchay, C. (2015). Selective deficits in episodic feeling of knowing in ageing : A novel use of the general knowledge task. *Acta Psychologica*, *157*, 85–92. doi:doi.org/10.1016/j.actpsy.2015.02.014

Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C., … Nadel, L. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. *Journal of Anatomy*, *207*(1), 35–66. doi:10.1111/j.1469-7580.2005.00421.x

Moulin, C. J. A., Conway, M. A., Thompson, R. G., James, N., & Jones, R. W. (2005). Disordered memory awareness: Recollective confabulation in two cases of persistent déjà vecu. *Neuropsychologia*, *43*(9), 1362–1378. doi:10.1016/j.neuropsychologia.2004.12.008

Moulin, C. J. A., Perfect, T. J., & Jones, R. W. (2000). The effects of repetition on allocation of study time and judgements of learning in Alzheimer's disease. *Neuropsychologia 38*, 748–756. doi:10.1016/S0028-3932(99)00142-6

Naveh-benjamin, M. (1987). Coding of spatial location information : An automatic process ? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *13*(4), 595–605. doi:10.1037/0278-7393.13.4.595

Naveh-Benjamin, M., Hussain, Z., Guez, J., & Bar-On, M. (2003). Adult age differences in episodic memory: Further support for an associative-deficit hypothesis. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(5), 826–37. doi:10.1037/0278-7393.29.5.826

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125–173. doi:10.1016/S0079-7421(08)60053-5

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. doi:10.1037/0033-2909.95.1.109

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "Delayed-JOL" effect. *Psychological Science*, *2*(4), 267–271. doi:10.1111/j.1467-9280.1991.tb00147.x

New, B., & Pallier, C. (2001a). Gougenheim 2.0. Retrieved from http://www.lexique.org/public/gougenheim.php

New, B., & Pallier, C. (2001b). Lexique Toolbox (software). Retrieved from http://www.lexique.org/toolbox/toolbox.pub/index.php

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: a new French lexical database. *Behavioral Research Methods, Instruments and Computers*, *36*, 516–524. doi:10.3758/BF03195598

Norman, E., Blakstad, O., Johnsen, Ø., Martinsen, S. K., Price, M. C., & Charles, L. (2016). The relationship between feelings-of-knowing and partial nnowledge for general knowledge questions, *7*(June), 1–7. doi:10.3389/fpsyg.2016.00996

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: a meta-analysis. *Psychology and Aging*, *23*(1), 104–118. doi:10.1037/0882-7974.23.1.104

Overgaard, M., & Sandberg, K. (2012). Kinds of access : different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1287–1296. doi:10.1098/rstb.2011.0425

Pedregosa, F., Varoquaux, G., Gramfort, A., Vincent, M., Bertrand, T., Grisel, O., … Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, J. W. (2007). PsychoPy--Psychophysics software in Python. *Journal of Neuroscience*

Methods, *162*(1–2), 8–13. doi:10.1016/j.jneumeth.2006.11.017

Peynircioğlu, Z. F., & Tekcan, A. İ. (2000). Feeling of knowing for translations of words. *Journal of Memory and Language*, *43*(1), 135–148. doi:10.1006/jmla.2000.2704

Poppenk, J., Köhler, S., & Moscovitch, M. (2010). Revisiting the novelty effect: when familiarity, not novelty, enhances memory. *Journal of Experimental Psychology-Learning Memory And Cognition*, *36*(5), 1321–1330. doi:10.1037/a0019900

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence Leak in Perceptual Decision Making. *Psychological Science*, *26*(11), 1664–80. doi:10.1177/0956797615595037

Rawson, K. A., & Van Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, *58*(3), 646–668. doi:10.1016/j.jml.2007.08.004

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology*, *19*(1), 90–138. doi:10.1016/0010-0285(87)90005-3

Reder, L. M., Liu, X. L., Keinath, A., & Popov, V. (2016). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic Bulletin & Review*, *23*(1), 271–7. doi:10.3758/s13423-015-0889-1

Reder, L. M., & Ritter, F. E. (1992). What determines initial Feeling of Knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(3), 435–451. doi:10.1037/0278-7393.18.3.435

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition* (Erlbaum, pp. 45–77). Hillsdale, NJ.

Reder, L. M., Victoria, L. W., Manelis, A., Oates, J. M., Dutcher, J. M., Bates, J. T., … Gyulai, F. (2013). Why it's easier to remember seeing a face we already know than One we don't: Preexisting memory representations facilitate memory formation. *Psychological Science*, *24*(3), 363–372. doi:10.1177/0956797612457396

Reggev, N., Zuckerman, M., & Maril, A. (2011). Are all judgments created equal? An fMRI study of semantic and episodic metamemory predictions. *Neuropsychologia*, *49*(5), 1332–42. doi:10.1016/j.neuropsychologia.2011.01.013

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology. General*, *137*(4), 615–25. doi:10.1037/a0013684

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin*, *137*(1), 131–48. doi:10.1037/a0021705

Ricks, T. R., & Wiley, J. (2009). The influence of domain knowledge on the functional capacity of working memory. *Journal of Memory and Language*, *61*(4), 519–537. doi:10.1016/j.jml.2009.07.007

Roberts, L. S., & Rhodes, G. (1989). Knowing your limits: Expertise and the feeling of knowing. *New Zealand Journal of Psychology*, *18*, 71–75.

Robin, J., Wynn, J., & Moscovitch, M. (2016). The spatial scaffold : The effects of spatial context on memory for events. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *42*(2), 308–315. doi:10.1037/xlm0000167

Sacher, M., Taconnat, L., & Souchay, C. (2009). Divided attention at encoding: effect on feeling-of-knowing. *Consciousness and Cognition*, *18*(3), 754–61. doi:10.1016/j.concog.2009.04.001

Saling, M. M., Berkovic, S. F., O’Shea, M. F., Kalnins, R. M., Darby, D. G., & Bladin, P. F. (1993). Lateralization of verbal memory and unilateral hippocampal sclerosis: evidence of task-specific effects. *Journal of Clinical and Experimental Neuropsychology*, *15*(4), 608–18. doi:10.1080/01688639308402582

Schacter, D. L. (1983). Feeling of knowing in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 39–54. doi:10.1037//0278-7393.9.1.39

Schreiber, T. A, & Nelson, D. L. (1998). The relation between feelings of knowing and the number of neighboring concepts linked to the test cue. *Memory & Cognition*, *26*(5), 869–83. doi:10.3758/BF03201170

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J.

(1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 3–29. doi:10.1037/0278-7393.23.1.3

Schwartz, B. L., Pillot, M., & Bacon, E. (2014). Contextual information influences the feeling of knowing in episodic memory. *Consciousness and Cognition*, *29*, 96–104. doi:10.1016/j.concog.2014.08.018

Schwartz, B. L., & Smith, S. M. (1997). The retrieval of related information influences tip-of-the-tongue states. *Journal of Memory and Language*, *86*(36), 68–86. doi:10.1006/jmla.1996.2471

Selmeczy, D., & Dobbins, I. G. (2014). Relating the content and confidence of recognition judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(1), 66–85. doi:10.1037/a0034059

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *Quarterly Journal of Experimental Psychology (2006)*, *65*(11), 2231–57. doi:10.1080/17470218.2012.685081

Shanks, L. L., & Serra, M. J. (2014). Domain familiarity as a cue for judgments of learning. *Psychonomic Bulletin & Review*, *21*(2), 445–53. doi:10.3758/s13423-013-0513-1

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition*, *9*, 313-23–6. doi:10.1006/ccog.2000.0450

Shimamura, A. P., & Squire, L. R. (1986). Memory and metamemory: a study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *12*(3), 452–60. doi:10.1037/0278-7393.12.3.452

Sidarus, N., & Haggard, P. (2016). Difficult action decisions reduce the sense of agency: A study using the Eriksen flanker task. *Acta Psychologica*, *166*(May), 1–11. doi:10.1016/j.actpsy.2016.03.003

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, *117*(1), 34–50. doi:10.1037/0096-3445.117.1.34

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *26*(1), 204–221. doi:10.1037//0278-7393

Souchay, C., Bacon, E., & Danion, J.-M. (2006). Metamemory in Schizophrenia: an exploration of the feeling-of-knowing state. *Journal of Clinical and Experimental Neuropsychology*, *28*(5), 828–40. doi:10.1080/13803390591000846

Souchay, C., Insingrini, M., Clarys, D., Taconnat, L., & Eustache, F. (2004). Executive functioning and judgment-of-learning versus feeling-of-knowing in older adults. *Experimental Brain Research*, *30*, 1–16. doi:10.1080/03610730490251478

Souchay, C., & Isingrini, M. (2012). Are feeling-of-knowing and judgment-of-learning different? Evidence from older adults. *Acta Psychologica*, *139*(3), 458–64. doi:10.1016/j.actpsy.2012.01.007

Souchay, C., Isingrini, M., & Gil, R. (2002). Alzheimer's disease and feeling-of-knowing in episodic memory. *Neuropsychologia*, *1442*, 1–11. doi:10.1016/S0028-3932(02)00075-1

Souchay, C., Moulin, C. J. A, Clarys, D., Taconnat, L., & Isingrini, M. (2007). Diminished episodic memory awareness in older adults: evidence from feeling-of-knowing and recollection. *Consciousness and Cognition*, *16*(4), 769–84. doi:10.1016/j.concog.2006.11.002

Staresina, B. P., & Davachi, L. (2009). Mind the gap: binding experiences across space and time in the human hippocampus. *Neuron*, *63*(2), 267–276. doi:10.1016/j.neuron.2009.06.024

Staresina, B. P., Gray, J. C., & Davachi, L. (2009). Event congruency enhances episodic memory encoding through semantic elaboration and relational binding. *Cerebral Cortex*, *19*(5), 1198–1207. doi:10.1093/cercor/bhn165

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*(6), 1213–20. doi:10.1037/0278-7393.31.6.1213

Starns, J. J., & Hicks, J. L. (2008). Context attributes in memory are bound to item information, but not to one another. *Psychonomic Bulletin & Review*, *15*(2), 309–314. doi:10.3758/PBR.15.2.309

Susser, J. A., Mulligan, N. W., & Besken, M. (2013). The effects of list composition and perceptual fluency on judgments of learning (JOLs). *Memory & Cognition*, *41*, 1000–11. doi:10.3758/s13421-013-0323-8

Thiede, K. W. (1996). The relative importance of anticipated test form at and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology*, *49*(4), 901–919. doi:10.1080/027249896392351

Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2012). An analysis of the determinants of the feeling of knowing. *Consciousness and Cognition*, *21*(4), 1681–94. doi:10.1016/j.concog.2012.09.005

Thompson, V. A, Turner, J. A P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–51. doi:10.1016/j.cognition.2012.09.012

Tiede, H. L., & Leboe, J. P. (2009). Illusions of competence for phonetically, orthographically, and semantically similar word pairs. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Expérimentale*, *63*(4), 294–302. doi:10.1037/a0015717

Trinkler, I., King, J. A., Spiers, H. J., & Burgess, N. (2006). Part or parcel? Contextual binding of events in episodic memory. In *Handbook of Binding and Memory: Perspectives from Cognitive Neuroscience* (pp. 53–83). doi:10.1093/acprof:oso/9780198529675.003.0003

Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press.

Tulving, E. (1985). Memory and Consciousness. *Canadian Psychology*, *26*(1), 1–12.

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: Role of the Hippocampus. *Hippocampus*, *8*, 198–204. doi:10.1002/(SICI)1098-1063(1998)8:3<198::AID-HIPO2>3.0.CO;2-G

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*, 381–391. doi:10.1016/S0022-5371(66)80048-8

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Rev*, *80*(5), 352–373. doi:10.1037/h0020071

Uncapher, M. R., Otten, L. J., & Rugg, M. D. (2006). Episodic encoding is more than the sum of its parts: an fMRI investigation of multifeatural contextual encoding. *Neuron*, *52*(3), 547–56. doi:10.1016/j.neuron.2006.08.011

Urquhart, J. A., & O'Connor, A. R. (2014). The awareness of novelty for strangely familiar words: a laboratory analogue of the déjà vu experience. *PeerJ*, *2*, e666. doi:10.7717/peerj.666

Van Overschelde, J. P., Rawson, K. A., Dunlosky, J., & Hunt, R. R. (2005). Distinctive processing underlies skilled memory. *Psychological Science*, *16*(5), 358–361. doi:10.1111/j.0956-7976.2005.01540.x

van Velzen, J. H. (2013). Students' explanations of their knowledge of learning processes. *Educational Studies*, *39*(1), 83–95. doi:10.1080/03055698.2012.671515

Vernon, D., & Usher, M. (2003). Dynamics of metacognitive judgments: Pre- and postretrieval mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 339–346. doi:10.1037/0278-7393.29.3.339

Voss, J. F., Vesonder, G. T., & Spilich, G. J. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior*, *19*, 651–667. doi:10.1016/S0022-5371(80)90343-6

Wandmacher, T., Ovchinnikova, E., & Alexandrov, T. (2008). Does latent semantic analysis reflect human associations? In *Proceedings of the Lexical Semantics Workshop et ESSLLI'08*. Hamburg, Germany.

Williams, C. C. (2010). Incidental and intentional visual memory : What memories are and are not affected by encoding tasks ? *Visual Cognition*, *18*(9), 1348–1367. doi:10.1080/13506285.2010.486280

Williams, H. L., Conway, M. A., & Moulin, C. J. A. (2013). Remembering and Knowing: Using another's subjective report to make inferences about memory strength and subjective experience. *Consciousness and Cognition*, *22*(2), 572–588. doi:10.1016/j.concog.2013.03.009

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. doi:10.1006/jmla.2002.2864

Zawadzka, K., & Higham, P. A. (2015). Judgments of learning index relative confidence , not subjective probability. *Memory and Cognition*, *43*, 1168–1179. doi:10.3758/s13421-015-0532-4

Zawadzka, K., & Higham, P. A. (2016). Recalibration effects in judgments of learning: A signal detection analysis. *Journal of Memory and Language*, *90*, 161–176. doi:10.1016/j.jml.2016.04.005