UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

School of Psychology

Bridging the Gap Between Learning and Memory

by

Gregory James Neil

Thesis for the degree of Doctor of Philosophy

August 2011

This page intentionally left blank

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

SCHOOL OF PSYCHOLOGY

Doctor of Philosophy

BRIDGING THE GAP BETWEEN LEARNING AND MEMORY

By Gregory James Neil

This thesis uses a fusion of recognition memory and implicit learning methods to investigate performance based on implicit learning. A series of experiments exposed participants to a study list composed of natural words that conform to a conjunctive rule-set involving the frequency and the concreteness of the words. Participants were asked either to identify words seen on the study list or to identify rule-consistent words. Across a variety of learning conditions signal-detection analyses revealed that participants used both the episodic status of the words (the episodic effect) and the structural status of the word (the structural effect) in making their decisions. Questionnaires indicated that participants could not verbalise the conjunctive rule-set. Increasing the number of repetitions of each word on the study list increased the magnitude of the episodic effect but not that of the structural effect. In addition, a classic strength-based mirror effect was found in which endorsements to words on the study list increased with repetitions but endorsements to both new rule-consistent and new rule-inconsistent words decreased. Discussion of recognition-memory models and a set of MINERVA simulations demonstrated that current recognition memory models cannot account for these results. Implicit learning theories also struggle to account for the invariance of the structural effect to repetitions. It is concluded that familiarity underlies both the structural effect and a portion of the episodic effect, but that the precursors of familiarity are different in each case with structural familiarity being insensitive to repetitions and episodic familiarity being sensitive to repetitions. Implications for recognition and implicit learning theories are discussed.

Table of Contents

List of Tables

List of Figures

Declaration of Authorship

I, Gregory James Neil declare that the thesis entitled "Bridging the Gap Between Learning and Memory" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission,

Signed:

Date:

Acknowledgements

I would like to thank my supervisor, Phil Higham, who gave me intellectual guidance and insightful comments throughout my PhD. I also thank him for having good humour when it was needed most.

This PhD was funded by the ESRC, whom I thank for the chance to do the PhD in the first place.

I would also like to thank the support staff at the University of Southampton for always being there with the answers to my (sometimes awkward) questions. Thanks must also go to all the people who participated in my experiments.

Thanks to my parents for supporting me through the years and for always encouraging me to do what makes me happy.

Last, but certainly not least, I thank my partner Flis for the constant support and advice. Without her I may not have made it through to the end!

Abbreviations

AG: Artificial grammar

AGL: Artificial grammar learning

ALT: Attention-likelihood theory

ANOVA: Analysis of variance

C: Criterion

CA: Common-abstract

CASK: Consciousness as king

CC: Common-concrete

CCRA: Common-concrete/rare-abstract

CFM: Calibrated familiarity model

CR: Correct rejection

CS: Strong criterion

CW: Weak criterion

EF: Episodic familiarity

FA: False alarm

FAR: False-alarm rate

GMM: Global-memory model

GSM: Global subjective memorability

HN: High new

HO: High old

HR: Hit Rate

LN: Low new

LO: Low old

M: Miss

MS: Milliseconds

RA: Rare-abstract

RC: Rare-concrete

RCCA: Rare-concrete/common-abstract

ROC: Receiver operating characteristic

RSI: Response-stimulus interval

RT: Reaction time

SAC: Source of activation confusion

SAM: Strength of activation model

SDT: Signal-detection theory

SF: Structural familiarity

SO: Strong old

SOC: Second-order conditional

SRT: Serial reaction time

UoS: University of Southampton

WO: Weak old

1 Chapter 1: Introduction

1.1 Overview

Initially this chapter reviews implicit learning and methods used to determine participants' awareness when completing implicit learning tasks, including signal-detection theory. Then, the specific claims of implicit learning research are reviewed before proceeding to take a look at the mirror effect in recognition memory. Finally, these strands are pulled together and some similarities are drawn between both literatures, setting the stage for an investigation of implicit learning which draws on recognition memory literature.

1.2 What is Implicit Learning?

Implicit learning has proven to be a slippery concept to define. The 1998 volume "The Handbook of Implicit Learning" devoted three chapters to defining it (Buchner & Wippich, 1998; Frensch, 1998; Stadler & Roediger, 1998) and even then, later chapters of the book took different perspectives on the definition of implicit learning. Although the definition may differ, there is no doubt that the concept of implicit learning is one worth exploring. Before discussing the definition of implicit learning in detail, it is worth reflecting on why implicit learning is of such interest.

1.2.1 Why research implicit learning?

Implicit learning is relevant to a broad spectrum of contexts. This section will discuss some of these contexts in order to demonstrate the value of research into implicit learning.

1.2.1.1 Mental health and brain injuries.

Implicit and explicit learning mechanisms are often thought to function in different ways. Some researchers have exploited this difference to investigate brain injuries and disorders. For instance, Knowlton, Ramus and Squire (1992) demonstrated that amnesiacs and non-amnesiacs performed equally when judging whether items matched a rule set. The same amnesiacs performed poorly when trying to remember whether they had seen the same items before. Knowlton et al. concluded that amnesiacs use implicit processes to drive performance, although Higham and Vokey (1994) disputed their findings. A similar result was found using word pairs (Musen & Squire, 1993) where implicit mechanisms accelerated amnesiacs' performance on a reading task to the same extent as controls.

Other examples of implicit learning research are the preservation of implicit learning in head trauma (Nissley & Schmitter-Edgecombe, 2002), distinguishing different types of obsessive-compulsive disorders by differences in function impairment (Rauch et al., 2007), deficits of implicit and explicit learning in schizophrenia (Pedersen et al., 2008), and adaptive strategies that schizophrenics use to overcome those deficits (Marvel et al., 2007).

1.2.1.2 Learning contexts.

Implicit-learning research has also focused on how people learn and how various circumstances affect this process. For instance, Kuhn and Dienes (2005) examined how people learn musical rules and demonstrated that implicit learning had greater sensitivity to these rules than did explicit learning. Several studies have examined how age affects the learning process. Pacton, Fayol and Perruchet (2005) found that children used implicitly learned rules of language to guide their performance on a task involving made-up words. Feeney, Howard and Howard (2002) found that middle-aged adults were slower and less accurate than younger adults in sequence learning. Bennett, Howard and Howard (2007) found that implicit learning declined in old age, although not as much as explicit learning.

Differing circumstances can affect implicit and explicit learning differently. For example, sleep deprivation does not affect how you use already implicitly learned information, but inhibits new implicit learning (Heuer & Klein, 2003). Miyawaki (2006) demonstrated differences in how explicit and implicit learning mechanisms react to changes in time between stimuli displays.

The characteristics of implicit learning compared to explicit learning are of interest in other learning contexts. Antony and Santhanam (2007) for instance demonstrated that a knowledge-based system designed to encourage implicit learning was effective in corporate skills training.

1.2.2 Defining implicit learning.

Implicit learning has resisted a consistent definition since A. S. Reber coined the phrase in a seminal paper on the topic (A. S. Reber, 1967). The problem is that not only is consensus on the definition of implicit learning hard to find, but most of the words that people use in definitions need further definition themselves. Diane Berry said of this in 1997:

> At this point in time it appears unlikely that we will come up with a definitive answer to the question of 'how implicit is implicit learning?' The answer will clearly depend on the criteria used to establish implicit learning, and at present there is no consensus on this.
> (D. C. Berry, 1997, p. 235)

Peter Frensch gathered together different definitions of implicit learning for the second chapter of The Handbook of Implicit Learning. He listed 15 definitions claiming that "literally dozens of definitions ... have been offered and continue to be offered in the literature" (Frensch, 1998, p. 51). Dimensions on which the 15 definitions differed were: intentionality of learning, awareness, conscious status, what is learned, intentionality of knowledge usage and automaticity. One way to define implicit learning is "learning that occurs at an unconscious level" – but then what constitutes unconscious? How can level of consciousness be measured? How can mode of learning be measured separately from the knowledge that springs from learning?

Many researchers sidestep the use of awareness in implicit learning by using the idea of intentionality. Such researchers encourage a switch between implicit and explicit learning by manipulating the wording of experimental instructions (e.g. D. C. Berry & Broadbent, 1988; Jimenez, Mendez, & Cleeremans, 1996; Song, Howard, & Howard, 2007). Implicit-learning instructions made no mention of learning anything about the stimuli. Explicit-learning instructions actively encouraged participants to learn something about the stimuli. Frensch (1998) suggested that an intentional versus unintentional distinction is more useful than an aware versus unaware distinction in an experimental setting. If participants adopt a rules-search strategy despite the lack of instructions to do so, an intentionality definition can break down and so experimental checks of learning strategy are needed when adopting this approach.

One of the most persistent questions about implicit learning concerns not *how implicit is implicit learning* but *what is learned in implicit learning*, *and how does it differ from what is learned in explicit learning*? When approaching this question, researchers are more likely to make claims of knowledge held without awareness (Kuhn & Dienes, 2005; Runger & Frensch, 2008; Tamayo & Frensch, 2007). Generally this research concerns the conscious state and contents of the knowledge. Measuring the conscious state of knowledge is a matter of enormous debate (see below).

For the purposes of this thesis, I will use *part* of the definition suggested by Frensch: Implicit learning is "the non-intentional, automatic acquisition of knowledge" (Frensch, 1998, p. 76). Frensch goes on to define the content of that knowledge. However, as the content of knowledge is highly debated, here "knowledge" will serve as a catch-all term to represent that *something* is learned. The next section deals with how to measure the conscious state of knowledge.

1.3 Measuring Implicit Learning and Implicit Knowledge

Most of the implicit learning experiments hinge around demonstrations of the knowledge gained as a result of implicit learning. For instance, Dienes and Scott (2005) showed participants a training list of strings of letters that were related to each other by a set of complex rules. They asked participants either to memorise items presented in a study phase (the implicit-learning condition) or actively search for rules (explicit-learning condition). Performance based on implicit knowledge was unaffected by study instructions. For explicit knowledge, performance was worse after rule-search instructions than after memorise instructions. But how can you be sure that measures of implicit and explicit knowledge are really measuring implicit and explicit knowledge?

1.3.1 Objective and subjective measures of awareness.

In order to classify knowledge as implicit, it is necessary to demonstrate that the experimental participant has no awareness of the information being used (Dienes & Perner, 1999). Awareness is difficult to measure. Most early experiments attempted to show lack of awareness by using "objective" measures. The logic of these studies is that if study phase conditions can be set such that participants identify stimuli only at chance levels, then they must be unaware of said stimuli. Cheesman and Merikle (1984) for instance displayed one of four words on each trial in a calibration phase. In the "unaware" condition they manipulated the display duration of each word until the participants correctly identified the word one in four times (i.e. chance-level responding). These display durations were then used to investigate the effects of priming assuming that the participants were not aware of the stimuli.

Jacoby (1991) advocated an alternative objective measure approach. If implicit knowledge is used automatically and explicit knowledge is used voluntarily (Buchner & Wippich, 1998) then implicit knowledge can operate in two ways. Either implicit knowledge works to the same end as explicit knowledge (facilitating) or implicit knowledge introduces inappropriate errors that need to be filtered out by conscious processes (interference). These

principles have seen wide-spread use since their conception, for instance in studies on the effects of motivation on conscious and unconscious influences (Visser & Merikle, 1999) and conscious and unconscious influences on artificial-grammar learning (Higham, Vokey, & Pritchard, 2000).

In contrast to objective measures, subjective-measure theories (Cheesman & Merikle, 1984; Dienes, 2008a) assume that awareness is a purely subjective state and can never be tapped by experimenter-defined objective measures. Cheesman and Merikle (1984) demonstrated that participants defined as unaware using an objective measure of awareness performed at chance on an experimental task. If instead unaware was defined as participants claiming to be unaware, then they found a level of above-chance performance that appeared to respond to manipulations differently compared to when participants claimed to be aware. Cheesman and Merikle concluded that participants classified as unaware by objective measures of awareness actually had no knowledge at all. In essence, the simplest measure of awareness is to ask. Participants could be reporting guessing when they actually have some level of conscious knowledge, but this point will be addressed later.

More recently, increasingly sophisticated subjective measures of awareness have been used. Dienes and Berry (Dienes, 2008a; Dienes & Berry, 1997) put a name to the two most common subjective report methods - "the guessing criterion" and the "zero-correlation criterion". These measures require an estimate of confidence for each individual trial from participants. For illustrative purposes, imagine a confidence rating from 0 to 100. By the guessing criterion, there is implicit knowledge when performance is above chance across responses on the basis of guessing (i.e., zero confidence responses). The zero-correlation criterion indicates implicit knowledge when above-chance performance is found in the absence of a correlation between confidence ratings and performance.

Shanks and St John (1994) illustrated several problems with measures of awareness. They specified two criteria that should be satisfied to conclude that implicit-learning mechanisms have been at work:

1. The Information Criterion – Proof is required that a change in performance is due to the knowledge that is being measured with tests of awareness. This means that the test of awareness must not measure something other than awareness, and the test performance must not be due to knowledge that is not being measured.
2. The Sensitivity Criterion – The test of consciousness must be sensitive to all relevant conscious knowledge. If it is not, then performance on the basis of

unmeasured conscious knowledge cannot be ruled out. In other words, the measure must be process pure.

Shanks and St John (1994) reviewed several experiments that claimed to show implicit learning, and concluded they all failed to meet these criteria. They went as far as saying that implicit, unconscious processes do not exist at all. The sensitivity criterion is particularly damaging to objective measures of awareness as it is virtually impossible to prove that *any* measure is not contaminated by other processes, let alone one intended to indicate awareness (Jacoby, 1991; Kunimoto, Miller, & Pashler, 2001; Merikle & Daneman, 1998; Reingold & Merikle, 1988)

These are difficult problems for measures of awareness, raising several questions. Does verbal report reflect just awareness and nothing else? Could a participant have been aware at the time of their primary judgement but forgotten by the time the test of awareness is administered? Is the information participants use to make their verbal report the same information that defines their awareness? Do we do things of which we have no conscious awareness?

Another criticism of measures of awareness is that the sensitivity of the measure can change depending on which rating scale you use. Tunney and Shanks (Tunney, 2005; Tunney & Shanks, 2003) found that a binary rating scale was more sensitive to explicit knowledge than a continuous rating scale, making the continuous rating scale capable of misrepresenting explicit knowledge as implicit. They suggested that continuous scales were more difficult for participants to use, although there was some evidence that participants may learn to use continuous scales with practice. Alternatively, the binary scale could be misrepresenting unconscious knowledge as conscious. Dienes (2008a) replicated the experiment, varying the difficulty of the task and found that for most conditions there was no sensitivity difference, other than one case where the continuous scale was more sensitive than the binary one.

Despite the problems with subjective measures, there is an array of evidence that indicates that they can be useful. "Just asking" is not always as imprecise as it sounds. Barnhard & Geraci (2008) used questionnaires to assess participants' states of awareness. Participants studied a list of words under deep or shallow learning conditions. After a five-minute distraction task, participants were given three-letter word-stems and asked to write the first word that came to mind beginning with those letters. In the 'study' condition, some of the word stems could be completed using the studied words. The 'no-study' condition contained no such stems. Participants were then given a questionnaire to see if they were

aware of using studied words to complete word stems. As word stems in the no-study condition could not be completed with studied words, this acted as a baseline condition of zero awareness. They found that the questionnaires were accurate in tracking predicted levels of awareness by depth and study conditions. If anything, they found that questionnaires were too conservative in assigning people to the unaware category. In the baseline condition the questionnaires assigned 83% of the no-study participants to the unaware category using a maximally conservative criterion. When they made the criterion slightly more liberal, the accuracy approached 100%. Thus, quite against Shanks and St Johns' (1994) assertion, it would appear that rather than conscious information contaminating measurements of unconscious processes, it is more likely that unconscious processes contaminate measurements of conscious performance.

J. Reed and Johnson (1998) made a similar argument wherein they accused Shanks and St John (1994) of adopting a "Consciousness as King" (CASK) approach. They argued that the sensitivity criterion assumes that explicit knowledge is a given; therefore the burden of proof is on showing that implicit knowledge exists. But there is no reason why implicit knowledge should not be a given, and so the sensitivity criterion could just as easily be turned around to require that tests of explicit knowledge be uncontaminated by implicit knowledge. A final nail in the coffin of the Shanks and St John criticisms comes from studies of what people actually do in day-to-day life. One such example is an experiment in which people were asked to describe how they go about catching a ball (N. Reed, McLeod, & Dienes, 2010). Despite that fact that participants could clearly catch the ball, few participants could describe the actual strategy that allowed them to position themselves correctly (which involved controlling the angle of gaze to the ball). Some of the incorrectly identified strategies were even given with very high confidence; in other words, confidence and accuracy were not related but objective performance was clearly above chance.

1.3.1.1 Bias.

There is one other problem with subjective measures – that of bias. There are several discussions of bias in the literature: Merikle and Daneman (1998) tackled bias in perceptual tests of awareness and Dienes (2004) discussed bias in implicit knowledge. The problem is that it is impossible to know what criterion individuals use to distinguish their own aware state from an unaware state. For example, consider an experiment in which a confidence rating is obtained by asking participants to respond either "guess", "relatively sure" and "very

sure" (Tamayo & Frensch, 2007). Individuals with a liberal criterion may choose guess only if they have no confidence whatsoever in their answer. An answer with any confidence at all would be assigned to either relatively sure or very sure – thus a liberal criterion will result in hardly any guesses and a large number of "sure" responses. A conservative individual may assign some answers in which they have a small amount of confidence to the guess category to ensure that the two 'sure' categories reflect only answers in which they have high confidence. Thus a conservative criterion will result in a large number of guesses. This bias can vary from individual to individual but also from task to task.

Dienes (2004) considered two types of bias. One type of bias springs from a disconnect between mental representations of knowing and the actual state of knowing. The other concerns the inaccurate mapping of a state of knowing to an experimental measure. Dienes concluded that the former is what allows implicit knowledge to exist at all and is what we want to study, whereas the latter is simply measurement error and needs to be eliminated. Cheesman and Merikle (1986) argued that bias can simply be ignored. Differences between conditions may be affected by bias but they are still interesting differences. If a manipulation has a different effect above and below a subjective threshold of awareness, then regardless of bias there is still an effect of interest to study. As Dienes (2004, p. 40) pointed out, "Progress will be made by theory development and maximising the usefulness of data for testing theory". If a theory makes predictions about how knowledge use will vary above and below a subjective threshold, which is backed up by data, then it is of interest regardless of bias.

One final way to deal with bias is to use a measurement method that is bias free, but that has the capability to measure bias itself.

1.3.2 Signal-detection theory.

Signal-detection theory (SDT) is a method originating in radio operations, where the operators had to distinguish a true signal from the general static they might receive. This idea has been applied in psychology for many years (see Macmillan & Creelman, 2005 for a good handbook). In memory research, you can define a signal as a word that a participant has been asked to remember from a study list (an "old" word) and noise as words not on the study list ("new" words). Take as an example an experiment where a participant is shown a study list of words to remember, and then given a test list containing new and old words. Participants must classify items on the test list as new or old items. In type-1 signal detection, when a participant sees a word on the test list they are assumed to experience a particular strength of

feeling (for instance a feeling of familiarity) which is called the strength of evidence. Words seen at study can trigger strength of evidence in a high range as they have been recently seen (the right hand distribution in Figure 1.1 – the target distribution) whilst new words trigger a lower range of strength of evidence (the left hand distribution – the distractor distribution). Participants are assumed to set a criterion (C in Figure 1.1) along the strength-of-evidence continuum. If a test item elicits strength of evidence above this criterion they respond old whilst if the strength of evidence is under the criterion they respond new. As the distributions overlap there is room for participants to misidentify which distribution a stimulus comes from. SDT classifies the possible responses into hits (participant says old when the word is in fact old), misses (participant says new when the word is old), false alarms (participant says old when the word is new) and correct rejections (participant says new when the word is new) - see Table 1.1.

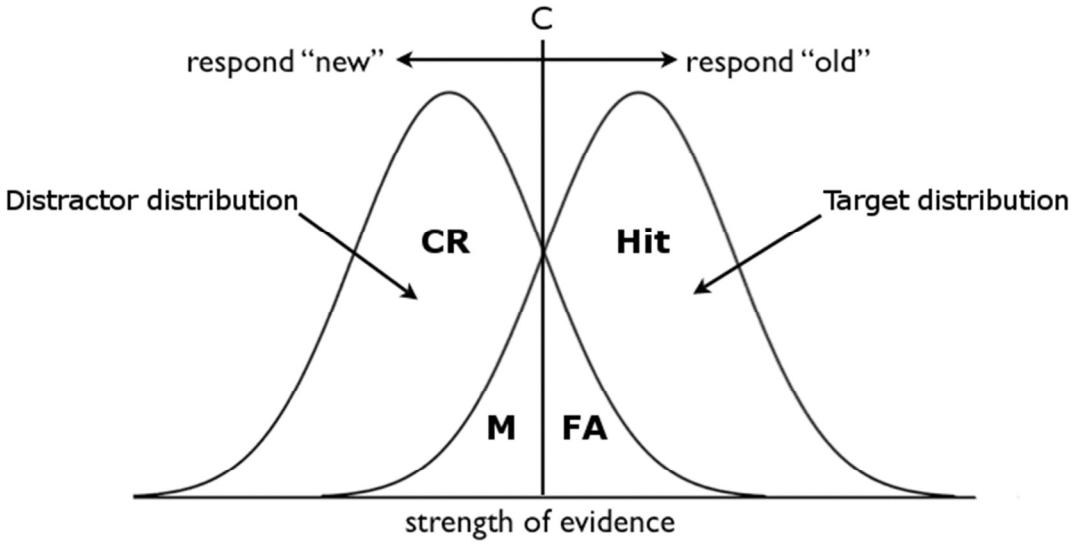

Figure 1.1. Type-1 Signal-detection theory. FA = False alarm, M = Miss, CR = Correct rejection, C = Criterion. The signal-detection measure *d'* estimates the distance between the peaks of the two distributions.

Table 1.1

Type-1 Classification in Signal-Detection Theory

Stimulus state	Participant Response	
	New	Old
New	Correct Rejection	False Alarm
Old	Miss	Hit

In SDT the hit rate (HR) and false-alarm rate (FAR) are used to estimate the distance between the two distributions. This is represented by a measure of discriminability – d'[1] is one such commonly used measure, although there are others. SDT also provides various methods to estimate the relative position of the criterion. These indicate how biased participants are to respond old or new regardless of the distance between the distributions. Thus SDT provides us with a way of measuring the ability to discriminate between stimuli, without bias affecting that measure. Examples of type-1 SDT use include showing that blindsight patients' vision is not like that of normal people (Azzopardi & Cowey, 1997) and investigating performance on recognition-memory tests (Glanzer & Adams, 1990).

Vitally for implicit learning research, Type-1 SDT does not address the issue of awareness directly. For that additional tests of awareness are needed, such as type-2 SDT. The application of type-2 SDT to issues of awareness is well demonstrated by Kunimoto et al. (2001) and reviewed by Galvin, Podd, Drga and Whitmore (2003). The x-axis in type-2 SDT represents confidence in accuracy instead of strength of evidence. The two distributions (see *Figure 1.2*) now represent a participant's correct and incorrect responses. The right hand distribution is composed of all of the instances where they call new items new and old items old, whereas the left hand distribution contains all old items called new and new items called old. The criterion now represents a level above which a person says "high" confidence and below which they say "low" confidence when asked about the accuracy of their responses. The hits, false alarms (FA), correct rejections and misses can be categorised as in Table 1.2.

[1] Pronounced dee prime

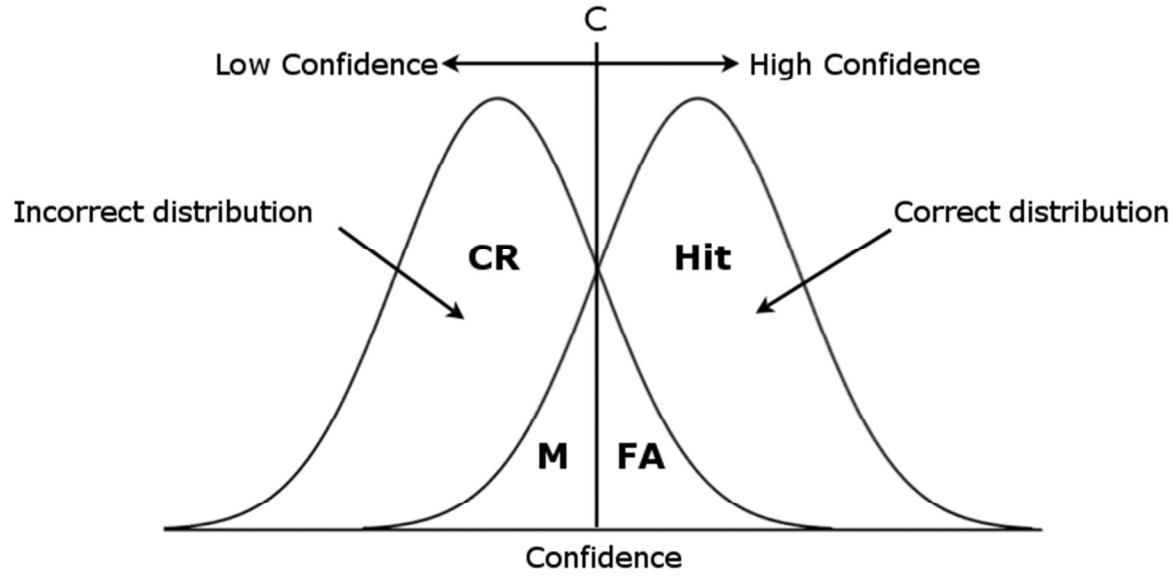

Figure 1.2. Type-2 Signal-detection theory. FA = False alarm, M = Miss, CR = Correct rejection. The signal-detection measure *d'* estimates the distance between the peaks of the two distributions.

Table 1.2

Type-2 Signal-Detection Classifications

Discrimination	Participant Confidence	
	High	Low
Correct	Hit	Miss
Incorrect	False Alarm	Correct Rejection

In type-2 SDT a hit occurs when the participant assigns high confidence to a correct response, a miss occurs when they assign low confidence to a correct response, a FA is high confidence assigned to an incorrect response and correct rejection is low confidence assigned to an incorrect response. Measures of discrimination such as *d'* can now be used to see how well a person can distinguish their own correct responses from their own incorrect responses, and bias measures reflect confidence bias. As a measure of awareness, it can now be said that a person has no awareness when their type-2 *d'* is zero – they cannot distinguish between their own correct and incorrect responses. Provided that their type-1 *d'* is above chance they have knowledge but no awareness. In essence the type-2 *d'* is a zero-correlation criterion measure.

Kunimoto et al. (2001) compared their type-2 SDT method with Cheesman and Merikle's (1986) experiment and demonstrated that Cheesman and Merikle's procedure was affected by bias, classifying some participants as unaware when they were in fact aware. Participants still performed above chance in the zone where the type-2 *d'* was zero (confidence and performance were unrelated). Perception without awareness was therefore demonstrated. Other studies have used type-2 SDT to analyse metacognitive aspects of cued recall (Higham, 2002; Higham & Tam, 2005), examination performance (Higham, 2007) and the generalisability of learned rules (Tunney & Shanks, 2003). Evans and Azzopardi (2007) criticised type-2 SDT by demonstrating that the type-2 *d'* can be sensitive to changes in bias. Despite their criticism however, Evans and Azzopardi conceded that Kunimoto et al.'s approach is useful when looking at thresholds of awareness.

The subject of SDT will be returned to in a later section concerning the mirror effect (section 1.6). The next section will present a review of studies about implicit learning. Following that, the mirror effect in recognition memory will be discussed, with the intention of comparing recognition and implicit learning research.

1.4 What is Learned in Implicit Learning?

Much of the research into implicit learning concerns what is learned when using an implicit mode of learning, how different processes of learning can be differentially affected by manipulations and what the properties are of the knowledge resulting from this learning. In the following sections evidence from the main paradigms are reviewed.

1.4.1 Artificial grammar.

1.4.1.1 Origins.

Artificial grammar (AG) experiments on implicit learning originated from an experiment by Miller (1958). Miller showed participants lists of nonsense words created by a rule set, and demonstrated that participants had an advantage when recalling rule-consistent words rather than rule-inconsistent words. He concluded that participants were coding parts of the words into memory in order to gain a recall advantage.

A. S. Reber (1967) ignited the last 40 years of AG research by developing this basic idea into the form that most AG experiments use today. He produced his materials using a Markovian finite-state network (see *Figure 1.3*). A series of consonant strings were produced by moving from node to node picking up the letters on the arrows, resulting in a string that

followed this specific rule set. Doing this several times resulted in a list of rule-consistent letter strings. For instance, every string must start with either a T or a P and end with either S or V. Although this example has two choices at each node, the actual number of choices per node can vary. Variations in length can be induced with loops and multiple paths. These strings are useful as they are free from previous associations and yet complex enough that they cannot be easily learned. The entire grammar cannot be inferred from studying one string. A string that follows the rules is called a grammatical string, and one that does not is called non-grammatical.

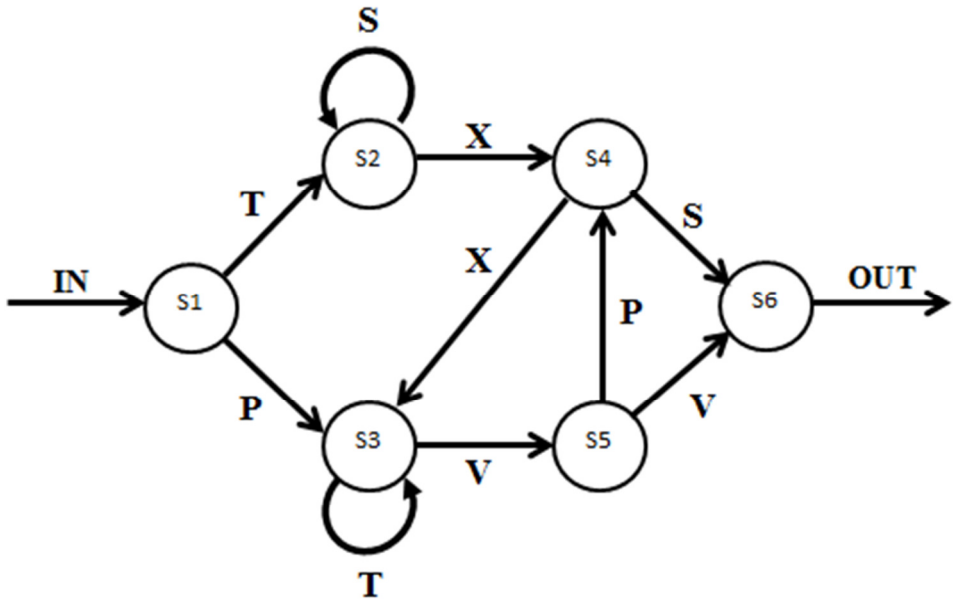

Figure 1.3. An artificial grammar network modified from A.S. Reber (1989)

Participants in A. S. Reber's experiment were asked to remember 20 grammatical strings for a future test in which they would have to recall them. Strings were grouped into four sets of five letters each and repeatedly displayed to participants until they could write down each set of five without looking at them. In the second phase of the experiment participants were told that the strings were created with a rule set, but not what the rule set was. A. S. Reber then showed them 88 new strings, some grammatical and some non-grammatical. The non-grammatical items had differing numbers of a variety of rules violations. Participants could identify the grammatical items at above-chance accuracy. They also correctly rejected non-grammatical strings more often if they contained more errors, and seemed to be particularly sensitive to violations at the beginning and ends of the strings. The number of "loops" in the string had no bearing on participants' judgements. The participants also could not explain what the rule set was when asked. A. S. Reber concluded

that people learned the rules by developing an internal model that mirrored the actual rule set, but that they do so by using implicit non-deliberate strategies.

This basic result has been replicated many times over the last four decades (Dienes, Broadbent, & Berry, 1991; Higham, 1997a; Knowlton & Squire, 1996; Lotz & Kinder, 2006; Perruchet & Pacteau, 1990; A. S. Reber, 1969; Scott & Dienes, 2008; Servan-Schreiber & Anderson, 1990; Tunney & Shanks, 2003; Van den Bos & Poletiek, 2008; Vokey & Higham, 2005; Whittlesea & Dorken, 1993). The fact that people can learn something that allows them to distinguish grammatical from non-grammatical strings is not disputed. The main debate that rages in AG literature is what people learn, and to what extent both the learning mode and the knowledge gained is implicit.

A.S. Reber characterised artificial-grammar learning (AGL) as not only developing an internal rule structure but as abstracting rules at a deeper level. He demonstrated that changing the letters used in the grammar from study phase to test phase resulted in no detriment to participants' grammaticality ratings (A. S. Reber, 1969). For instance if the letters used in the grammar at study were MVTXS, you can swap M for Z, V for F and so on. Provided the underlying grammar was the same, participants performed above chance. When the underlying grammar changed then performance dropped to chance. A. S. Reber concluded that participants were not learning aspects of the exact stimuli, but were abstracting the underlying rule set. No participant could explain the rules and so it was concluded that the knowledge was implicit. In a later review of his and others work, A. S. Reber (1989) further characterised implicit learning as being idiosyncratic. In three conditions he used a normal study-phase, but showed people the actual schematic of the grammar either at the beginning, in the middle or at the end of the training. Seeing the schematic at the beginning enhanced participants' learning of the grammar, but seeing it later was detrimental. Reber suggested that participants were forming abstract rules based on the grammar but representing those rules in a unique way. Seeing the schematic resulted in a shift from their idiosyncratic representation to a representation based on the schematic. When participants were asked to look for rules at training they actually performed worse than those who did not look for rules. A. S. Reber concluded that the primary learning mode was implicit.

Not everyone agreed with A. S. Reber's (1989) claims. Brody (1989) criticised A. S. Reber for not doing enough to show the state of awareness of the participants in his experiments, so conclusions about the unconscious state of the knowledge could not be drawn. Other disagreements disputed the claim that abstract rules were learned. For instance

it is possible that there was more than one process at work, such as a comparison of the test item to all items in memory – a "chorus of instances" - in order to make a decision (Vokey & Brooks, 1992). A similar theory is that it is not the stimuli that are processed or their attributes, but the experience as a whole (Whittlesea & Masson, 2005).

1.4.1.2 Single-process models of artificial grammar learning.

Single-process accounts of AGL focus on just one aspect of the stimuli being used to drive performance. One such account is that only the surface structures of the stimuli explain performance in AG tasks, such as the letters used to make the strings and how participants encode them. Dulany, Carlson and Dewey (1984) and Perruchet and Pacteau (1990) advocated the idea that micro rules about the surface structure are responsible for AGL. Dulany et al. demonstrated that although participants could not explain the overall grammar, they *could* underline viable parts of the strings, such as letter pairs (bigrams). Simulated rule-sets from the underlined parts of the strings were enough to reconstruct participants' grammaticality performance. Perruchet and Pacteau extended this idea and compared groups trained on normal study-strings with those who purely saw valid bigrams in training. Performance between the two groups at test was equivalent. Items that were non-grammatical due to non-permissible bigrams were easier to reject than items that were non-grammatical due to permissible bigrams in the wrong place. The conclusion was that performance was entirely due to consciously held bigram knowledge. Dienes et al. (1991) also found that partial strings were enough to explain classification performance. Jamieson and Mewhort (2009a) similarly suggested that AG performance could be due to knowledge of surface features. They used a recognition-memory model called MINERVA 2 (Hintzman, 1984, 1986) to simulate performance in an AG experiment. They found that the model could simulate AGL, likely because a rule-based grammar imposes constraints as to which letters would appear in which conditions, and MINERVA can learn the frequencies with which each letter appeared in each location. As MINERVA is a single-process model, this suggests that AGL relies on a single process. Jamieson, Holmes and Mewhort (2010) also criticised the Knowlton et al. (1992) assertion that amnesia evidence suggested an implicit/explicit dual-process explanation. They simulated the original Knowlton et al. data with a modified MINERVA 2 model and demonstrated that the task differences could be explained by amnesiacs having lower data integrity than non-amnesiacs. The MINERVA 2 model could replicate the data pattern rendering a dual-process account unnecessary.

Servan-Schreiber and Anderson (1990) took the idea a step further. They suggested that participants could learn strings as "chunks" which they then use to construct and evaluate higher-level strings. For instance, someone might learn the string TTXVPXVS as three chunks of TTX, VP and XVS. They called each of these chunks "words" (TTX) and theorised that participants fit them together into "phrases" (TTX VP) and that the overall strings were made up of phrases fit together (so TTX VP combines with XVS to create TTXVPXVS). They theorised that grammaticality judgements were based on how many new chunks would have to be coded to cope with that string – so TTX VP GGG only needs one new chunk coding and so would be judged grammatical but XTT PS GGG needs all new chunks and would be rejected. They showed that participants were more likely to accept grammatical items if chunks and phrases learned in the training phase were preserved. Participants also rejected non-grammatical items when the phrase structure was violated.

Channon et al. (2002) provided support for chunk strength by using a different method of creating grammatical strings. This allowed them to vary chunk strength to a greater extent than had previously been possible. Chunk strength, introduced by Knowlton and Squire (1994), is the number of times that chunks in a test item appear in the actual training items. The higher the chunk strength, the more times a participant will have seen that chunk before. Channon et al. demonstrated that responding at test was not affected by grammaticality, but participants did tend to rate the higher chunk-strength items as grammatical more often than the lower chunk-strength items. The zero-correlation and guessing criteria indicated this knowledge to be implicit.

Other studies have developed broader concepts. Jamieson and Mewhort (2005) borrowed the ideas of grammatical, local and organisational redundancy from information theory. Grammatical redundancy refers to the degree of repeated information in the grammar as a whole. Local redundancy refers to the amount of redundant information in a specific string as compared to others the grammar might create and is measured by the number of alternate stimuli that you can create from a particular stimulus. For instance RBBB can only be rearranged a small number of ways therefore it has high local redundancy. Organisational redundancy measures the amount by which the representation of a stimulus could be reduced. For instance RBRBRBRB has high organisational redundancy - if RB was coded as 1 the strings length can be halved to 1111. Jamieson and Mewhort compared grammatical and ungrammatical stimuli and found that the increased organisational redundancy of stimuli constrained by a rule set was responsible for performance in their task and not the stimuli or the underlying rules.

Brooks and Vokey (1991) cautioned against interpreting data such as A. S. Reber's (1967) original letter-transfer results as evidence for abstract rule learning. They demonstrated that transfer-like performance can be obtained on the basis of similarities between strings rather than knowledge of a rule set. For instance, if a string in the original letter set has a triplet (i.e. MXVVVM) then participants are likely to endorse a transfer string with a different triplet (i.e. BDCCCB). Gomez, Gerken and Schvaneveldt (2000) implicated repetition structure as the main component of AG transfer performance. They replicated A. S. Reber's letter changing experiment using tightly controlled stimuli. For the transfer test they distinguished between strings with letter repetitions (i.e. MMRRDD) and strings that had only sequential relations (i.e. MRDABC). Gomez et al. found that for transfer strings above-chance performance was entirely due to strings with repetitions: strings with only sequential relations produced chance performance. They concluded that repetition structure is the main basis of AGL. Scott and Dienes (2010) demonstrated that such information was likely to be used without awareness, as performance on the basis of repetition structure was only found when participants thought they were guessing. They also found that participants' ratings of familiarity predicted their grammaticality responses, suggesting that grammaticality judgements were made on the basis of familiarity. However, Tunney (2010) found that participants' ratings of similarity predicted grammaticality performance and confidence ratings. One major feature in each case was the repetition structure of the stimuli, although other stimuli features also contributed. It remains to be seen if Tunney's similarity rating and Scott and Dienes' familiarity rating reflect the same underlying mechanism.

1.4.1.3 Multiple-process accounts of artificial grammar learning.

Despite the evidence that only surface structure is learned in AGL, a wealth of evidence exists that deeper learning takes place via multiple processes. Two processes are usually cited but the identity of these two processes differs from theory to theory. Some theories cite familiarity and recollection in a similar manner to recognition-memory literature, whilst others focus on the explicitness or implicitness of the knowledge. Others use dual process to refer to the fact that two different aspects of the stimuli are learned. For instance, Whittlesea and Dorken (1993) proposed a flexible episodic-based processing system where an implicit-learning mechanism is sensitive to the task requirements. By varying the method of coding at study and the task demands of the test, they demonstrated that participants can be induced to learn the structure of individual items, remember the experience of processing

items in a particular way (such as reading or hearing an item) or learn abstract elements of the general structure. The key finding was that abstractions did not happen when not required by the task. In experiments where bigrams are sufficient to perform the task, abstractions would not occur. Poznanski and Tzelgov (2010) similarly warn that task requirements can affect performance. They used intentional, incidental and automatic conditions in both a study phase *and* in the test phase. They found that different combinations of conditions at study and test produced different patterns of performance, even finding that automatic learning led to chance performance when tested with intentional or incidental conditions but above-chance performance when tested with automatic conditions. They concluded that task conditions should be carefully set, and that many different aspects of AG strings contribute to performance. In terms of familiarity and recollection, Tunney (2007) demonstrated that participants would pick both "remember" and "familiar" options when presented with new rule-consistent stimuli at test. Tunney concluded that both recollection and familiarity processes contributed to grammaticality judgements, even when the test items had not been previously seen.

Knowlton and Squire (1996) claimed that explicit knowledge of chunks is epiphenomenal and although chunks aid performance, implicit rule-abstraction is the dominant factor in AGL. They conducted an experiment with controls and amnesiacs. For both groups, grammatical items were endorsed regardless of their chunk strength but non-grammatical items were only judged grammatical if their chunk strength was high. As the amnesiac group performed at chance levels on a recognition test, Knowlton and Squire concluded that both chunk strength and rule abstraction were learned and used implicitly.

Previous results involving chunk strength may be due to the number of exemplar items shown at study. Meulemans and Van der Linden (1997) varied the length of study lists and found that when the lists were short, participants appeared to use chunk strength to make grammaticality judgements. When more exemplars were available, chunk strength did not contribute to performance. They theorised that a rule-abstraction mechanism was used in this case. As participants were unable to generate valid strings on their own, Meulemans and Van der Linden concluded that the knowledge was implicit.

Higham (1997a) also claimed that chunk strength alone does not explain performance on AG tasks. He used stimuli that were either balanced or not across chunk strength and similarity (similar strings had 1 letter changed and non-similar items had more than 1 letter changed). When chunk strength was balanced, classification performance due to grammaticality was above chance. Manipulating how pronounceable the words were

changed performance even when chunk strength was held constant. In a separate study, Higham (1997b) demonstrated that distraction affected performance differently when comparing chunk balanced to chunk unbalanced stimuli. Grammaticality performance increased from balanced to unbalanced stimuli even when distracted. However similarity-based performance existed only for unbalanced chunk-strength stimuli and was eliminated by distraction. Higham concluded that a dual-process model is required to explain these results, with similarity primarily consisting of an explicit recall process and grammaticality using a primarily implicit familiarity-based process. Crucially, models citing chunk strength would also have to employ such dual-process explanations to explain the different effects of chunk strength across distraction. Knowlton and Squire's claim that chunk strength is purely implicit could not explain these results. Further experiments also led to similar dual-process conclusions (Higham et al., 2000; Vokey & Higham, 2004). Vokey and Brooks (1992) found similarity and grammaticality effects, and theorised that both effects may be under the control of different processes. They demonstrated that if you individuate items then it becomes harder to use the global similarity of test stimuli to study stimuli to make decisions. However grammaticality effects stayed constant when items were individuated, suggesting the underlying rules of the grammar were learned.

In contrast to Gomez et al.'s (2000) theory that repetition structure is responsible for learning in AG tasks, Tunney and Altmann (2001) claimed that transfer on the basis of statistical learning can occur. They demonstrated that if there is a large amount of noise in the statistical relationships between letters in training items then learning based on these relationships is disrupted. For instance, either G or B can follow A with equal probability and the study items reflect that. Noise exists in the study items if none of the test items have any instances of B following A, as learning that relationship will not help performance. With noise, transfer occurred only by repetition structure. Without noise, transfer occurred using both statistical relationship and repetition. Tunney and Altmann claimed that implicit learning in AG uses both statistical rule learning and exemplar comparison and that Gomez et al.'s results were due to noisy grammars.

Taking a different approach, Lotz and Kinder (2006) used regression to predict performance in a normal AG task and a transfer task. They concluded that in the normal AG task participants used repetition information, but only in the non-transfer condition did they use information about chunks. Vokey and Higham (2005) also demonstrated that transfer works using repetition information. They used stimuli that were matched to the study-list stimuli such that test stimuli were either similar or non-similar to study items. The data were

analysed by how many repetitions there were in the stimuli (e.g., ABCDE had no repetitions, AABCD had one repetition and AABBC had more than one). The letters used to represent the grammar from study to test were changed. In the consistent change condition, each study letter was mapped onto a new letter for the whole test phase. In the random condition, every individual test item used a new letter set. They found that with both one and many repetitions, but not with zero repetitions, participants endorsed similar items more than non-similar items. Letter change affected the use of grammaticality but not the use of similarity. They concluded that grammaticality performance was in part based on surface features, but that similarity-based performance worked though comparison to a "chorus of instances" in which participants experienced a feeling of familiarity by retrieving multiple study items that were similar to the test item.

Kinder, Shanks, Cock and Tunney (2003) applied a perspective from memory literature to AGL, claiming that the information learned via implicit learning can sometimes be deployed flexibly. They demonstrated that the feeling of processing an item quickly (fluency) influenced peoples' performance in AG tasks. Resolving some of the stimuli on screen faster than others created a false feeling of fluency. When the test stimuli did not include any items that were on the study list (i.e. old items), fluency increased grammaticality judgements, but with old items on the test list there was no effect of fluency. Kinder et al. attributed this to the presence of old words inducing a switch from fluency-based judgement to recollection-based judgement. Buchner (1994) also investigated the effects of fluency on grammaticality judgements. In a task where stimuli were initially covered by a black square and then slowly revealed, he found that grammatical items were identified more quickly than non-grammatical items. The identification time did not predict a subsequent grammaticality judgement. By relating fluency to familiarity, Buchner concluded that familiarity was not the sole basis of grammaticality judgements.

Interference has also been used to support a dual-process model with explicit and implicit knowledge as the two processes. Tunney and Bezzina (2007) and Tamayo and Frensch (2007) both demonstrated that a time interval had differential effects on measures equated to implicit and explicit-based performance. Performance due to explicit knowledge decreased with greater time intervals but performance due to implicit knowledge was untouched. Tamayo & Frensch induced the same pattern by introducing a time interval between study and test.

Van den Bos and Poletiek (2008) claimed that previous experiments that favoured single-process accounts did so because the stimuli prevented implicit learning from operating.

They varied the dependency length of the rule set – i.e. how many letters had to be remembered in order to learn relationships. For instance, an eight-letter string where the seventh letter depends on the four preceding letters would have a high dependency length. Participants in the implicit-learning condition displayed better grammaticality performance than those in the explicit-learning condition. Only implicit learning was impaired by increasing the dependency length. Knowledge of first-order dependencies (i.e. bigrams) was learned in all conditions. If some of the single-process studies used grammars with high dependency length, then implicit learning would not have operated and bigrams would account for all performance.

One recent model that integrates these findings is that implicit learning results in implicit knowledge which influences performance through familiarity, whilst explicit learning results in implicit and explicit knowledge (Scott & Dienes, 2008). The model states that when learning implicitly, participants establish a mean familiarity that they expect to feel when looking at test stimuli, and use variations from the mean to judge grammaticality. Small variations are only barely felt and are not judged as predictive. These small variations appear unconscious by the guessing criterion. Larger variations are experienced more strongly and are assigned confidence that registers as conscious application of knowledge. In both cases, participants cannot verbalise the grammar or rule set, but measures of familiarity predict performance. In explicit learning conditions familiarity alone did not account for performance, indicating the influence of a second, explicit, process. Although the pre-cursors of implicit knowledge were not consciously known, participants felt a feeling of familiarity based upon those pre-cursors. Such a result demonstrates how knowledge which participants cannot verbalise can influence performance.

1.4.1.4 Artificial grammar summary.

Although AG experiments have produced valuable evidence about implicit and explicit learning and knowledge, no firm conclusion has been reached on the basis of this evidence. Are single- or dual-process accounts needed to explain performance? Is learning of simple surface structures enough to explain performance or is something else needed? What surface structures do we learn – repetition, chunks or similarity? And how might a deeper mechanism work – by rule abstraction or by comparison with a chorus of instances?

As task demands have been shown to vary what is learned, these sorts of questions still require extensive research to answer. Findings that implicit performance can be

explained by familiarity can help to focus this research and shed light on the underlying mechanisms. A definitive answer to what is learned in AG tasks is still tantalisingly out of reach. A second source of evidence about implicit learning comes from a different paradigm – serial reaction time (SRT) tasks.

1.4.2 Serial reaction time tasks.

1.4.2.1 Origins.

Since 1987, SRT experiments have become popular for investigating implicit and explicit learning and knowledge. Nissen and Bullemer (1987) are credited with the method that developed into the standard SRT. SRT experiments typically present participants with a screen containing four areas for stimuli (such as asterisks) to appear. The possible locations are made obvious by dividing the screen into four quadrants. Participants are informed that they are part of a reaction time experiment and must press the button that is linked to the location as quickly as possible, whilst being careful not to make mistakes. The stimuli are displayed in the locations according to a pattern. For example if the top left is 1, top right is 2, bottom left is 3 and bottom right is 4 then the sequence might be 1234123412341234. The sequence would be repeated many times – SRT tasks usually have more than a thousand stimulus displays, arranged into blocks. Participants were not informed of the existence of the rules generating the sequence in early versions of the task, but later versions can involve rule-search instructions. Amount of learning is measured by the reduction in reaction time to the sequence over the course of the experiment.

Modern SRT tasks use a "transfer" block to control increased reaction speed that occurs as a result of practice. In this block, a semi-random sequence (random other than the constraint that the same location is not used twice in a row) is used rather than a sequenced one. Reaction times (RT) from the sequenced blocks preceding and succeeding the transfer blocks are then compared with RTs from the transfer block. The difference is the degree to which participants have learned the sequence. Tests of consciousness in the SRT task usually take the form of a generate task. Assuming above-chance performance, if participants can generate the sequence themselves they are said to be explicitly aware of the sequence rules. Failure to generate the sequence is taken to mean they have implicit knowledge of the sequence. A discussion of problems with SRT consciousness measures will be reserved until later (see section 1.4.2.3).

SRT researchers take a perspective similar to AG researchers, investigating what is learned and the conscious state of the knowledge resulting from learning. The learning mode is often assumed to be implicit. For example, Heuer and Klein (2003) tested two groups of participants on two different sequences over a period of days. Both groups performed the learning of the first sequence normally and then immediately performed a test of learning. That night, one group was deprived of sleep. The next day they were tested again on the first sequence, then learned a second sequence which they were immediately tested on, with a final delayed test administered later in the day. The results showed that sleep deprivation affected new learning, but not the use of knowledge previously obtained under non-sleep-deprived learning. The use of implicit knowledge was assumed, and in fact any participant showing high performance was removed from the analysis under the assumption that they had developed explicit knowledge of the sequence.

The SRT paradigm has been used extensively with a variety of special populations including amnesia, trichotillomania and obsessive compulsive disorder (Goldman et al., 2008; Rauch et al., 2007), schizophrenia (Marvel et al., 2007; Pedersen et al., 2008), and the elderly (Bennett et al., 2007)

1.4.2.2 Single-process models of serial reaction time learning.

Like AG strings, the sequences in SRT experiments can contain chunks. J. Reed and Johnson (1994) described the "surface features" of a repeating sequence:

1. *Transition frequency* – the number of times that each possible transition appears in a sequence. A sequence of 1234 would have the transitions 12 23 and 34 for instance.
2. *Reversal frequency* – how many times the sequence goes back on itself. So 121 is a reversal – the stimulus appears in the top left, then top right, then back to the top left.
3. *Rate of full coverage* – the average number of trials needed until each zone has been used at least once.
4. *Rate of complete transition usage* – the average number of trials needed until each possible transition has been used.

J. Reed and Johnson (1994) argued that any of these simple features could be especially salient in a structured sequence compared to a random sequence. People learn these features which are correlated with the underlying structure, rather than learning the

actual structure. Because these simple features are salient, they could also appear in tests of explicit knowledge, possibly making it look like a participant has explicit knowledge of a sequence when they actually have explicit knowledge of the surface features. For instance, if a sequence has the reversal 121 then that section may be learned as a triplet because of the reversal, or if a sequence has a transition that it is highly frequent (the "12" part of the 6 item sequence 123412) then the transition may be learned as a bigram or trigram. When J. Reed and Johnson compared sequences holding these features constant to those that varied them, the features were indeed learned.

Later SRT experiments avoided these confounds by using complicated structures and controlling for simple features (e.g. Norman, Price, Duff, & Mentzoni, 2007; Rowland & Shanks, 2006; Song, Howard, & Howard, 2008). One common form is a "second-order conditional" (SOC) sequence, in which every third number can be determined from knowing the preceding two. For example two triplets might be 124 and 132, where various different digits can follow the initial digit, but the third digit is defined entirely by the two before it. The complexities of these sequences produce fewer salient transitions, controlling for simple features. Another version of the SRT task uses probabilistic relationships. For example, given the pair 12, the third digit will be 3 60% of the time and 4 40% of the time. Learning is indexed by the difference between the high and low probability trials (e.g. Jimenez, Vaquero, & Lupianez, 2006; Remillard, 2008). SOC and probabilistic trials are more difficult to learn than first order conditional or deterministic sequences and thus likely to encourage implicit knowledge use.

Although some researchers have stated that implicit learning is limited to learning SOC sequences (Jimenez et al., 1996), there is evidence to suggest that in the right circumstances implicit learning can result in knowledge of third and even fourth order sequences. Remillard and Clark (2001) used first, second and third order transitions and tested awareness with a verbal self-report questionnaire. They found evidence of learning in all three transition conditions and the verbal report showed knowledge was implicit for all three conditions. Additionally, the learning decreased from first to second to third order transitions. Remillard and Clark argued that this reflected a dependency-length limitation: as the number of items that have to be remembered increases, the efficiency of implicit learning decreases. In later work, Remillard (2008) extended this using a probabilistic design. By adjusting the probabilities to make higher-order conditional relationships slightly easier to learn, he demonstrated that up to fourth-order conditional relationships could be learned. Impressively, the relationships need not be adjacent - there can be lag between parts of the

sequence and the relationship was still learned. As there was no evidence of awareness from participants Remillard concluded that the learning and knowledge was all implicit and based on the statistical dependencies inherent in the structures rather than simple surface characteristics.

Most single-process explanations of SRT learning do not involve chunks. J. Reed and Johnson (1994) found learning even when all simple features had been controlled. Jimenez (2008) criticised one of the few studies to suggest chunks as the main mechanism for SRT learning. When a sequence was arranged in obvious triplets (123 321 for instance) participants had slower reaction times at the beginning of a new triplet, and faster reaction times for the second and third item of the triplet (Koch & Hoffmann, 2000). Jimenez demonstrated that this tendency was present in participants from the very beginning. The pattern of results could be explained by having to reverse direction, or change the hand that was being used to respond. In other words, properties of the motor control of the fingers produced these patterns, not chunk learning.

Motor control and learning are often implicated as part of performance in SRT tasks (Abrahamse, van der Lubbe, & Verwey, 2008; Song et al., 2008). Despite this, single-process explanations of SRT tend to conceptualise implicit and explicit knowledge as part of a single continuum rather than all learning being motor learning. French, Buchner and Lin (1994) argued that only an associative mechanism is needed to explain the results of an interference experiment. Participants had to attend to tones at the same time as performing an SRT task. For each trial, either a high or low frequency tone was played and participants had to count the high frequency tones. The tones were played either at the same time as the stimulus display, 350 milliseconds (ms) afterwards or 700 ms afterwards. They found that the immediate and 700 ms display participants evidenced more learning than the 350 ms participants. The 350 ms lag was believed to interfere with the grouping of stimuli in short-term memory, making associations hard to form. Frensch et al. concluded that an associative mechanism was the central form of implicit learning.

Jimenez et al. (2006) argued that performance in SRT tasks can be explained through a single-process model in which implicit knowledge is a weaker form of explicit knowledge. Participants were given intentional (explicit) learning instructions or unintentional (implicit) learning instructions. An opposition-logic generation task was used to check the conscious state of the knowledge (in the inclusion condition, participants are asked to recreate the sequences; in the exclusion condition, they are asked to avoid recreating the sequence). In a random-sequence block the occasional structured-sequence run was included. Participants

given explicit instructions noticed that the block looked random and abandoned using their knowledge of the sequence, thereby showing only chance performance in this block. The implicit instruction group performed above chance on the sequenced parts, as they did not intentionally stop using sequence knowledge. If there were dual processes of implicit and explicit knowledge at work, the explicit instructions participants should still have performed above chance in the random block, which they did not. Jimenez et al. concluded that there is a single system in which representations that are too weak to be used voluntarily or strategically are implicit knowledge and stronger representations that can be used in this fashion are explicit knowledge. Fu, Fu and Dienes (2008) drew a similar conclusion. They demonstrated that participants could be induced to show more explicit knowledge (as indexed by an opposition generation test) when they were offered rewards or allowed more practice. Using this result, they argued that implicit knowledge is weak knowledge that does not support awareness of the content of that knowledge. Jamieson and Mewhort (2009b) also detailed a single-process explanation by demonstrating that MINERVA could simulate SRT data. Each trial was stored in memory as a single vector of simple features, which meant that entries accrue throughout an experiment. Each vector included information about the preceding response. Performance could be explained by comparing the current trial with the vectors in memory. For SRT data, memory is probed using the previous trial to search for the current response. As more vectors accumulate in memory, the current response is obtained faster, mimicking the decrease in RT seen in SRT experiments.

Shanks, Rowland and Ranger (2005) used an interference experiment to support their view on single-process models. Interference was instantiated by presenting task stimuli in one of four different symbols and asking participants to count the occurrence of a particular symbol. They found that RTs were higher with interference than in a standard SRT and that an opposition logic generation task showed knowledge to be explicit. They argued that it is therefore questionable that implicit learning results in non-verbalisable knowledge at all. Rowland and Shanks (2006) demonstrated that having symbols on the screen imposed a 'filter cost' in that their presence reduced RTs, but did not reduce the amount of learning found. As there was no dissociable influence here they again claimed that implicit learning results in explicit knowledge.

1.4.2.3 Multiple-process accounts of serial reaction time learning.

Despite some support for single-process accounts, much SRT research seems to favour a dual-process explanation, although different specific processes are often implicated. Miyawaki (2006) suggested a dual-learning explanation. The pattern of intervals between a participant pressing a key and the presentation of the next stimulus (the response-stimulus interval or RSI) was manipulated as were task instructions. With instructions to simply respond to the stimuli (implicit learning) Miyawaki found that changing the RSI pattern from long-short-long to short-long-short greatly impaired performance. However, under rules-search instructions (explicit learning) the RSI change had a much smaller effect. Miyawaki concluded that there were dual processes at work, whereby implicit learning was sensitive to constraints and groups imposed by environmental factors (the RSIs) whereas explicit learning was capable of linking across environmental factors. In this sense, implicit learning is sensitive to simple associative relationships, whereas the explicit mode can use the implicitly learned information and link it together in more complicated sequences. Stadler (1995) demonstrated a similar result, where asking participants to count tones or varying the RSI reduced learning under implicit conditions. He concluded that the ability of implicit learning to organise the stimuli was disrupted, preventing analysis of dependencies between items. Norman et al. (2007) hypothesised different graduations of consciousness in SRT performance, where performance could be obtained using explicit knowledge with an implicit basis such as an explicit feeling of familiarity based on an implicit understanding of the rules.

Abrahamse et al. (2008) used a variant SRT task to suggest a different kind of dual-process model. He used a SOC sequence, but administered the experiment in two different ways. One group trained on the sequence as normal with a screen in front of them, learning the sequence by sight and response (the vis-tac group). The other group had four speakers taped to their fingers, and they trained on the sequence using the vibrations of the speakers (the tac-vis group). Each group, having completed the standard experiment, were given a short test on the other stimuli set (i.e. the speaker group responded on the basis of visual stimuli and vice versa). They found that the tac-vis group displayed perfect performance transfer, but the vis-tac group did not. Via a generate test, the vis-tac group demonstrated a low level of awareness but the tac-vis group demonstrated no awareness at all. Two modes of learning and response were implicated. The tactile condition resulted in pure implicit motor learning of which the participants were not aware. The visual condition resulted in motor learning (which could be transferred) as well as stimulus-specific learning about the

sequence, which was partly explicit but also aided by eye movement effects. Song et al. (2008) controlled for eye movement and found that there were still two identifiable processes, concluding as Abrahamse et al. did that both motor- and perceptual-based learning contribute to performance.

The most common demonstrations of dual-process models in the SRT domain rely on interference, distraction, or cues to show differential effects on modes of learning and knowledge use. Song et al. (2007) conducted a SRT in which their sequence was interspersed with random elements. If S is part of the structured sequence and R is a random element the stimuli followed the pattern SRSRSRSR. In most of the SRT blocks, the random elements were cued with black squares and the sequenced elements were cued with red squares. Implicit learning participants were not told the meaning of the squares, but explicit learning participants were. All participants had three blocks in which all the stimuli were presented with black squares regardless of structure. As participants with explicit knowledge have been found to abandon its use when it appears no longer relevant (Jimenez et al., 2006), Song et al. used the all black blocks to probe for the effects of implicit learning. They found that both groups demonstrated equivalent implicit-learning performance, but that the group with explicit instructions had further performance advantages. The explicit learning did not change accuracy, but did improve RTs. As measured by a suite of different tests, the implicit-learning condition only resulted in implicit knowledge whereas the explicit-learning condition resulted in both explicit and implicit knowledge. Song et al. argued for a dual-process account in which implicit learning occurs independently of explicit learning. Additionally explicit knowledge use interfered with implicit knowledge use initially, but that tendency faded in later blocks. Song et al. (2008) also found that giving intentional-learning participants instructions to stop searching for rules or continue searching for rules had different effects depending on the learning mechanism. In participants who learned in the usual SRT way by responding to stimuli, the transfer instructions had no effect. In participants who learned by watching the sequence, continue-search instructions led to chance performance but the stop-search instructions resulted in equivalent performance to the respond group. The tests of awareness used indicated that all knowledge was implicit. They concluded that both rules-based learning and motor learning occurred, and that the implicit rules-based knowledge can be interfered with by task demands such as cues or intentional search instructions.

Abrahamse and Verwey (2008) changed the test context to investigate dual-process models. They used an alternating sequence in which every other element followed a

sequence and the rest were random. Rather than displaying stimuli in four corners of the screen the stimuli were displayed in four rectangles. When these rectangles were changed to triangles performance suffered, mainly in the sequenced elements, but also slightly in the random elements. Two possible explanations were advocated: either the shapes were coded as part of the sequence, or changing the shapes resulted in participants re-evaluating the task causing explicit knowledge to interfere with the expression of implicit knowledge. Both explanations implicate a dual-process model involving motor learning and dependency-based sequence learning.

Evidence that people learn the dependencies between stimuli has also been provided by SRT like tasks. Goschke & Bolte (2007) used a six item repeating sequence, but instead of using asterisks appearing in different positions they used categories to instantiate the sequence. There were four categories of items, and many different items that could appear as part of each category. They used a generation task and a recognition task to test the conscious state of the knowledge. Overall, participants learned the category sequence, even when they were not told the categories. Participants appeared to abstract from the surface characteristics in order to learn the underlying pattern. Although this could be interpreted as learning the underlying abstract rule-set or as exemplar-based learning, it is certainly different from motor learning and indicates that people use processes other than pure motor learning in SRT tasks. Perlman and Tzelgov (2006) used a sequence instantiated with colour words written in different colour inks (i.e. green written in yellow) to show that learning of a sequence occurred in all combinations of automatic and non-automatic learning and retrieval.

Some studies indicate that evidence from SRT experiments that rely on knowledge measurement should be approached with caution. In Song et al.'s (2008) experiment it appeared that both motor and perceptual learning was occurring, but how they interacted was unclear. Song et al. believed that it was possible for perceptual learning to limit the expression of motor learning. This could complicate the measurement of motor learning, unless the perceptual element is carefully controlled. Runger and Frensch (2008) demonstrated that changing the demands of the task unexpectedly can change the level of verbalisable knowledge reported. They suggested that only unexpected events that do not require a heavy use of resources result in a "fruitful" search for sequence knowledge, and that this happens only with highly salient sequences. However, their hypothesis was not tested on SOC sequences. Norman et al. (2007) dealt a damaging blow to studies that claim explicit knowledge on the basis of generation tests. In an SRT, task they administered a standard generation test alongside a modified generation test that could only be completed using

explicit knowledge. Participants in an implicit learning performed well at the standard generation test but at chance levels in the modified generation test. Norman et al. concluded that results attributed to explicit knowledge in standard SRT awareness tests could be obtained on the basis of implicit knowledge.

1.4.2.4 Serial reaction time task summary.

SRT experiments have done much to distinguish between implicit and explicit modes of learning. However, what is learned in these modes remains undecided. Motor learning appears to be implicated as *part* of the learning but is not responsible for all of the performance. There is no definitive evidence to favour statistical abstraction or rule learning over learning of surface features or the use of an associative process (which amounts to a neural network version of the chorus-of-instances approach).

The added problems with measurement of awareness inherent in SRT experiments render conclusions difficult. If SRT explicit awareness measures can be completed on the basis of implicit knowledge then results showing disassociations cannot be trusted. As some SRT experiments also show that implicit learning or knowledge can be disrupted depending on sequence properties or explicit knowledge use, it is clear that future SRT researchers should be careful how they design their stimuli and of what conclusions they draw.

1.4.3 Evidence from other experiments.

Although AG and SRT experiments have provided insight into implicit and explicit learning and knowledge, valuable evidence has been obtained using other methods. D.C. Berry and Broadbent (1988) conducted a much-cited study using a dynamic control task. A user input numbers into a computer on the basis of which the computer generated responses. The aim was for participants to maintain the computer-generated number at a particular level. The task had different underlying algorithms: salient (the response depended on the number just entered plus or minus one) or non-salient (the response depended on a number entered on a *previous* trial plus or minus one). The appearance of the task also differed between controlling a person's emotions, the output of a sugar factory, passengers on a bus or passengers on a train. D. C. Berry and Broadbent demonstrated that all participants learned to control the task, but those training on a non-salient algorithm did so with no verbalisable knowledge. If a hint was given, performance only improved in the salient groups. In the non-salient groups, performance was *worse* after the hint – similar to results indicating

interference of implicit knowledge use by explicit knowledge. Hints also impaired transfer between conceptually similar tasks (i.e. from bus to train tasks). Positive transfer occurred between conceptually similar conditions, but not when participants were explicitly told they were similar. Again, an instruction encouraging explicit use of knowledge prevented transfer.

Bright and Burton (1994) also cited saliency as a factor in a clock-based experiment. All the clocks in a training phase followed the rule of displaying times between six and twelve. They manipulated type of clock seen at study and test and method of studying. Participants preferred clocks that followed the rule and types of clocks that matched the study clocks. Increased preferences were shown when the study task was conducive to coding information about the time displayed. Bright and Burton interpreted the different study conditions and clock faces as increasing the saliency of either the rule sets or the surface features. No participant could verbalise the rules used. Bright and Burton argued that implicit learning abstracted rules at a deep level, and in addition used simple surface features to enhance performance. Didierjean (2007) also implicated saliency as facilitating superior learning for rules involving even numbers compared to those with odd numbers.

In a complex problem-solving task, participants developed no verbalisable knowledge of three possible strategies to solve a task. However, analysis of performance at the task indicated that there was a point past which the task was performed as if such knowledge were held (P. J. Reber & Kotovsky, 1997). Protocol analysis was used to ensure that participants did not have knowledge at the time that they forgot later, but talking during the task interfered with participants' performance. It was concluded that implicit learning requires working memory to be successful.

In a twist on the AG style experiment, Higham and Brooks (1997) tested the hypothesis that participants can implicitly pick up on an experimenter's choice of stimuli for study lists. Items in a study phase were created using a conjunctive rule-set – the lists were composed of either common verbs *and* rare nouns or common nouns *and* rare verbs. The test phase used words seen at study (old), words not seen at study that were rule consistent (new consistent, NC) and words not seen at study that were rule inconsistent (new inconsistent, NI). Either a classification task or a recognition task was administered. Participants' sensitivity to the rule set was measured by comparing endorsement rates for NC words with those for NI words. In the classification task, people were sensitive to the difference between old and NC words indicating that they used recollection of old words to aid in classification judgements. In the recognition task, participants endorsed NC words more often than NI

words, indicating that participants were influenced by rule set. There were no instructions to look for rules, and the post-experimental questionnaire filtered out participants with verbalisable knowledge of the rules. It was concluded that participants used an explicit exemplar-comparison knowledge mode and an implicit rule-based knowledge mode for both tasks.

Foerde, Poldrack and Knowlton (2007) advocated a dual-process model in an experiment where participants were asked to learn which ice-cream flavour a succession of Mr Potato Heads would prefer. Learning occurred when participants were distracted and when they were not. Verbalisable knowledge only manifested when participants were not distracted. Distraction after learning prevented explicit knowledge from being used whilst leaving implicit knowledge use intact. Sun, Zhang, Slusarz and Mathews (2007) took a computational approach. They used different computational modes to simulate previous experiments. The only computational account that produced the correct data was one where implicit learning is initially used to assess information and then the output of this mode is used explicitly to test hypotheses.

Pacton, Perruchet, Fatol and Cleeremans (2001) provided support for a single-process learning mechanism by conducting experiments on several different year groups in a French school. Transfer sometimes incurs a performance reduction. Advocates of rules-abstraction learning mechanisms believe this indicates an incompletely learned rule structure. In theory, if the rule-learning mechanism were given enough time then the performance reduction should vanish. Pacton et al. tested five year groups on a transfer task that mirrored their real-life grammar, and found a consistent performance drop across all years. Network models using a similarity-based learning system were enough to explain their data. Kuhn and Dienes (2005) on the other hand demonstrated that in learning rules instantiated with music, participants learned non-adjacent dependencies. However, this information only appeared in indirect tests and in implicit subjective-knowledge conditions. In explicit-knowledge conditions participants appeared to learn only surface characteristics – in this case musical chunks.

Visual-search paradigms have demonstrated some possible confounds that could appear in implicit-learning experiments. Jiang and Song (2005) trained participants on a task in which participants had to find a letter L hidden amongst letter Ts. When participants were trained using materials of just one colour, performance gains were transferable. If they were trained on materials of two different colours, then training from one colour set did not transfer to the other. They suggested that the implicit-learning system uses the difference in

colour to eliminate possible strategies. Task differences between experiments may thus result in differential impediments of implicit factors. Rausei, Makovski and Jiang (2007) used similar materials to investigate attention. They concluded that the use of cues without awareness in a visual-search task required a minimum amount of attention. Above that minimum, additional attention did not further aid performance. Turk-Browne, Junge and Scholl (2005) demonstrated that only attended stimuli resulted in sequence learning but that no reportable knowledge of the sequence was developed. They suggested a two-stage process in which attention is required for learning to take place, but that the learning occurs without awareness.

1.5 A Synthesis of the Different Findings

All of the research seems to converge on the conclusion that some form of unintentional learning can occur. However, there is still some disagreement on what is actually learned. Often the evidence from each paradigm can be fairly specific. For instance, motor-learning conclusions from SRT experiments do not generalise to AG tasks, and generation task problems in SRT tasks can be circumvented in AG tasks by the use of trial-by-trial subjective measures.

Evidence from both SRT and AG tasks supports transfer from one set of characters to another, but the basis of this transfer is not clear. There is equal weight in the literature for both transfer on the basis of surface characteristics, and transfer on the basis of deeper knowledge. The question of whether there are multiple types of knowledge, or systems, is not yet resolved.

There is hope for a solution though. The application of methodologies such as SDT and subjective measures of awareness will help to disambiguate some of these issues. Theories that state that implicit knowledge results in explicit feelings help to explain how something we have no verbalisable knowledge of can influence our behaviour. There is another source of evidence relevant to the implicit/explicit debate which remains underused – recognition memory literature.

1.6 Recognition Memory and the Mirror Effect

The methods and terminology used in recognition literature are similar to those used in implicit-learning literature and there have been occasional experiments that take advantage of this fact (e.g. Higham & Brooks, 1997). Dual versus single processes is a recurring theme in the recognition memory literature (Glanzer & Adams, 1985; Jacoby & Dallas, 1981;

Mandler, 1980) as well as the implicit learning literature. Rather than being implicit and explicit knowledge, the dual processes are often cited as being familiarity and recollection (Reder et al., 2000). In experiments where old/new judgements must be made, models such as the source of activation confusion (SAC) model tend to typify recollection as a conscious memory of having studied a word and familiarity as a feeling the word has been seen before but with no specific memory of source. Familiarity has parallels with implicit knowledge as familiarity is often thought to be an automatic process, whereas recollection has parallels with explicit knowledge as both are thought to require conscious effort to activate.

Since recognition-memory experiments use similar methods to implicit-learning experiments, techniques developed in recognition memory should be useful in the implicit-learning domain. In this section I will focus on one important phenomenon from recognition literature - the "mirror effect".

1.6.1 What is the mirror effect?

Research into the mirror effect was primarily stimulated by two papers in the 1970's – Glanzer and Bowles (1976) and Brown, Lewis and Monk (1977). Imagine a simple recognition-memory experiment: participants are shown a list of words at study and then given an old/new test phase where they have to indicate which words were on the study list and which were not. The words on the study list differ along one dimension, for example how rare each word is in the English language (referred to as word frequency), and as such are classified into rare and common words. If these stimuli classes differ in their memorability then the higher memorability class will have better recognition performance than the lower memorability class. This difference will be due to a higher acceptance rate for old words (a higher HR) and a lower false acceptance rate for new words (a lower FAR). This is called the mirror effect. The word-frequency mirror effect occurs if lower frequency words have a higher HR and a lower FAR than the higher frequency words.

The mirror effect is both pervasive and reliable. Glanzer and Adams (1985) reviewed literature in which the effect has been found to occur for word frequency, concreteness, meaningfulness, type of stimuli (words versus pictures) and several other factors. The effect has been demonstrated where the different types of stimuli are present on the same training list (within-list manipulation) or on different training lists (between-list manipulations). Signal-detection experiments have also demonstrated it by increasing study-item strength via additional repetitions of training items (Stretch & Wixted, 1998). In one study, it was even

demonstrated that the mirror effect can operate in a computer-game task with many distracting features, prompting the authors to state that the mirror effect is highly generalisable (Ozubko & Joordens, 2008).

1.6.2 Explanations of the mirror effect.

Explanations of the mirror effect are usually presented in a signal-detection framework (section 1.3.2). There are two main classes of models: one seeks to explain all mirror effects in terms of differentiation of distribution locations whilst the second suggests different mechanisms for different types of mirror effects. Figure 1.4 shows one example of the differentiation model favoured by Glanzer and associates (Glanzer & Adams, 1990; Glanzer & Bowles, 1976; Glanzer, Kim, & Adams, 1998). Participants are shown a study list and then undertake a recognition test. Just like in standard SDT models, when a participant views an old word they experience a high strength of evidence whereas a new word results in a lower strength of evidence. New and old words can result in different ranges of strength of evidence, and so mapping these on a strength-of-evidence scale results in two overlapping distributions (what the strength of evidence actually is will be returned to later, as it is a matter of debate). In a recognition experiment, participants are thought to set a decision criterion. If a test item has strength of evidence that is equal to or higher than the criterion, participants say "old". Otherwise, they say "new". Responses can be classified exactly as in the SDT model (section 1.3.2). Rather than two distributions, Glanzer's model uses four and it is the ordering of these distributions along the strength-of-evidence axis that creates the mirror effect. Take, for example, the word-frequency mirror effect with high- and low-frequency words – the four distributions would be called high new (HN), high old (HO), low new (LN) and low old (LO). Low-frequency words are discriminated better than are high-frequency words, so the distance between the means of the LN and LO distributions must be greater than the distance between that of the HN and HO means. In order for the low-frequency words to have a lower FAR than the high-frequency words, the mean of LN must be below that of HN. Conversely, in order for low-frequency words to have a higher HR than high-frequency words, the mean of LO must be higher than that of HO. Thus the overall order of the distribution means is LN < HN < HO < LO. The mirror effect is a result of this pattern of distributions – low-frequency words have a higher HR and a lower FAR than high-frequency words. In Glanzer's model, participants actually use a likelihood ratio (see section 1.6.2.1) in order to make their decision. The use of a likelihood ratio means that even

strength-based mirror effects are explained by differentiation of distributions – strong stimuli are not only stronger in memory but also easier to differentiate from lures, whereas weak stimuli are held less strongly in memory and are more easily confused with lures. Thus strong items have higher HRs than weak items because of their memory strength, and lower FARs than weak items due to being more distinct from lures.

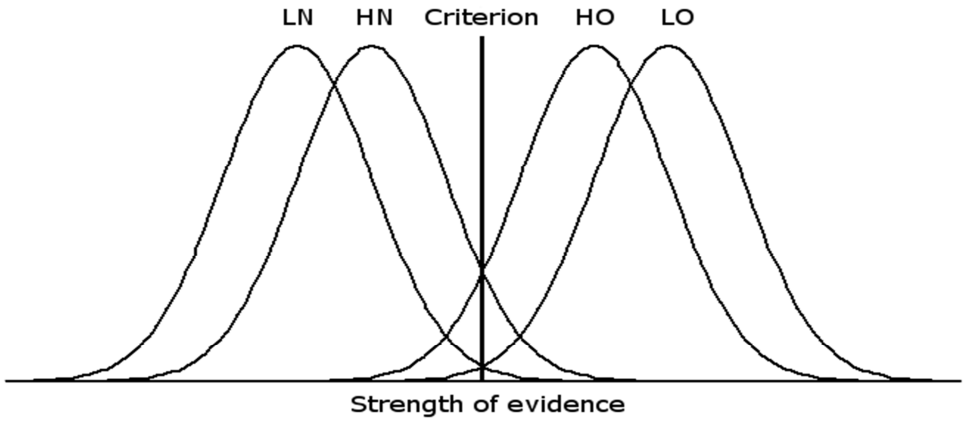

Figure 1.4. Different distribution explanation of the mirror effect. LN = Low new, HN = High new, HO = High old and LO = Low old.

The second class of models (e.g. Stretch & Wixted, 1998) is based on a criterion shift rather than different distributions, again using a SDT framework (see *Figure 1.5* below). This class of models accepts that word-frequency (and similar) mirror effects operate through different distributions which exist due to different levels of pre-existing familiarity for the high- and low-frequency stimuli. However, strength-based mirror effects are explained with a criterion shift. Study strength will be used as an example. Study items are presented either once (weak) or five times (strong). By the criterion-shift model, when a participant considers a stimulus more memorable they require a higher strength of evidence to respond "old" than to a less memorable word. Thus a conservative criterion is adopted for the memorable strong words and a liberal criterion is adopted for the less memorable weak words. Since there is no difference in strength of evidence for new words by study strength, the more conservative criterion for strong words results in fewer FAs than does the liberal criterion for weak words. As strong words are actually more memorable than weak words, the mean of the strong old distribution is higher up the strength-of-evidence scale than is the weak old distribution. The shift from the weak to the strong distribution is greater than the shift from the weak to strong criterion, resulting in a higher HR for the strong words compared to the weak words. Thus

the pattern of increased HR and decreased FAR from weak to strong words is produced. Brown et al. (1977) also suggested that people may shift their criterion in this way.

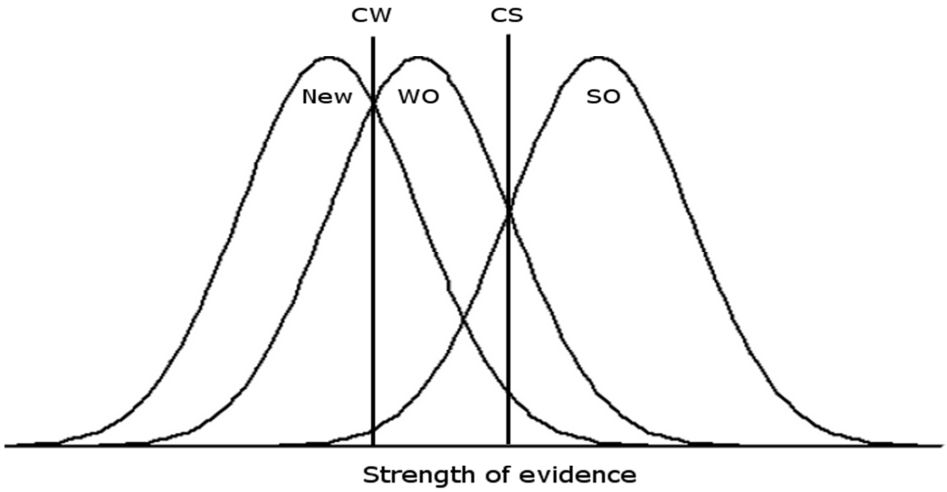

Figure 1.5. Criterion shift explanation of the mirror effect. CS = Strong criterion, CW = Weak criterion, SO = Strong old and WO = Weak old.

1.6.2.1 Likelihood.

Although all models utilise a strength-of-evidence dimension in one way or another, how it is used and what it represents is a source of disagreement. One class of models that reject the idea of a criterion-shift explanation of the mirror effect do so by assuming that the evidence dimension represents a likelihood ratio (Glanzer & Adams, 1990; McClelland & Chappell, 1998). A likelihood ratio compares the probability that an old word could produce the observed strength of evidence to the probability that a new word could produce the observed strength of evidence. This creates a ratio which is then compared to the stationary likelihood criterion. The assumption is that participants will say old when the ratio is greater than 1:1 – or, in other words, participants say old when it is more likely that the strength of evidence was produced by an old than a new word. Different distributions will produce different profiles of likelihood ratios, but the criterion stays at 1:1. For instance, Glanzer and Adams (1990) attention-likelihood theory (ALT) hypothesises that when a word is seen at study a number of features are marked in memory for that word. The distributions of high- and low-frequency words are separated because viewing high- and low-frequency words demands different levels of attention. More attention is paid to low-frequency words than to

high-frequency words. This differential attention leads to more features being marked for low- than high-frequency words and an increased distance between new and old distributions for low-frequency words. The increased distance means that the ratios tend towards the extreme ratios for the low-frequency words but not for the high-frequency words. This makes the mapping of likelihood ratio on to decisions more accurate for low- than for high-frequency words resulting in more hits and fewer FAs. Glanzer, Adams and Iverson (1991) extended the model further to include forgetting.

Likelihood explanations and ALT in particular are sometimes criticised for requiring too much computational effort on the part of the participant (e.g. Hirshman, 1995). Additionally, Stretch and Wixted (1998) used a variety of study strength and frequency-based manipulations to demonstrate that participants do not use target and lure information to make decisions about stimuli and so a likelihood ratio would be difficult to calculate. Although they dismiss likelihood explanations, Stretch and Wixted (1998) suggest that frequency-based mirror effects follow the multiple-distribution model and strength-based mirror effects follow the criterion-shift model. More recent likelihood models such as REM (Shiffrin & Steyvers, 1997) can include criterion shifts but do not actually cite these as explanations of effects, favouring instead explanations based around odds ratios.

1.6.2.2 Criterion shifts.

Criterion shifts have been demonstrated many times when a between-list manipulation is used. Cary & Reder (2003) demonstrated a between-list mirror effect by manipulating the length of the study lists. Both Criss (2006) and Stretch and Wixted (1998) showed between-list effects by using one study list that included each word once (the weak condition) and another study list that included each word five times (the strong condition).

Within-list criterion changes have been harder to find. Wixted and associates (Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998) have suggested that participants set their criterion once at the beginning of a test block and then are unwilling to expend the effort required for an item-by-item criterion shift. Higham, Perfect and Bruno (2009) made it obvious that there were two different classes of stimuli for a within-list manipulation of study strength by using all living words (e.g. duck) for the weak condition and all non-living words (e.g. table) for the strong condition for half the participants, with the association reversed for the other half. They found no evidence that participants shifted their criterion on an item-by-item basis at test. In a follow up study, Bruno, Higham and Perfect (2009) found that

participants can be induced to use such cues at test to shift their criterion, but only when participants believed the task to lead to generally low memorability conditions, such as when stimuli were briefly presented. They dubbed this the global subjective memorability (GSM) hypothesis. The GSM hypothesis implies that previous findings of a lack of criterion shift for within-list manipulations was due to the fact that participants judged the study phase to result in a high level of memorability. Singer (2009) also found a within-list criterion shift by requiring semantic ratings at study rather than using remember-the-stimuli instructions.

Brown and Steyvers (2005) have suggested a different explanation for the lack of within-list criterion shifts. They administered 10 blocks of 100 trials to participants and asked them to decide if each presented string was a real word or a non-word within a time constraint. The easy set had non-words that had three letters changed from a real word equivalent whereas the hard set had non-words with only one letter changed. They changed from the easy to hard set at various different points in the experiment – sometimes changing between test blocks and sometimes within test blocks. They found that the criterion could not shift instantly, even when participants were specifically told to do so. It took a lag of around 14 trials to shift the criterion to the appropriate place for the difficulty of materials. Since most studies using within-list designs display test stimuli randomly, this may explain why criterion shifts have not been found – participants do not have enough time with one stimulus type to adjust their criterion. The lack of ability to intentionally control item-to-item criterion shifts could be compared with the lack of intentionality of implicit knowledge use. Greene and Thapar (1994) even suggested that since participants do not seem to be able to pick up on stimuli differences the mirror effect may be due to unconscious information.

1.6.2.3 Similarities with implicit-learning research.

As in implicit-learning research, mirror-effect research has many references to the possibility of dual processes. Often these references identify the processes as recollection and familiarity (e.g. Joordens & Hockley, 2000). The SAC model also uses processes of recollection and familiarity to explain the mirror effect (Reder, Angstadt, Cary, Erickson, & Ayers, 2002; Reder et al., 2000). Recollection is said to stem from a specific memory of seeing a stimuli in the experimental context whereas familiarity is experienced as a feeling with no assigned source. The FA portion of the mirror effect is explained by differences in the pre-experimental familiarity of stimuli – with high-frequency words having a higher base-line familiarity than low-frequency words. The HR portion, on the other hand, is due to the

fact that the activation level of the experimental context node, created by an exposure to a stimulus at study, is influenced by the number of concepts otherwise associated with that stimulus. Low-frequency stimuli have fewer pre-experimental associations than high-frequency stimuli, resulting in a more efficient recollection process for low-frequency words. Thus a stimulus will either be recalled or not, and if it is not recalled then familiarity is used to make the decision. In general, decisions on the basis of recollection can be explained by the participants as having seen the stimulus at study. Decisions on the basis of familiarity are explained by a vague feeling of having seen the stimulus somewhere before but not being able to explain where. In this way, familiarity is similar to implicit knowledge – participants can make accurate decisions on the basis of both familiarity and implicit knowledge but they cannot explain what exactly it is that produces the feelings.

Models such as SAC assume that recollection is an all or nothing process – either it happens, or it does not. If it does not, then decisions are made on the basis of familiarity. However, several objections to that model have been raised in the literature (Higham & Vokey, 2004; Wixted, 2007; Wixted & Stretch, 2004). Some recognition research separates recollection from familiarity by the use of remember and know attributions. If recollection is all or nothing and the remember category represents recollection then there should be no new words inappropriately identified as old words (FAs) in the remember category. However, most recognition experiments do find FAs in the remember category, and have no mechanism to explain them. Wixted and Stretch (2004) suggested a signal-detection interpretation which naturally predicted such patterns. Evidence based on familiarity and recollection is summed to produce the strength of evidence represented by the axis in the SDT model. Remember and know judgements are made on the basis of the experienced strength of evidence, with remember being chosen at high strengths of evidence and know being chosen at lower levels of strength of evidence. As some amount of the new distribution overlaps with the placement of the remember criterion, FAs are a natural prediction of this model. As with measures of implicit and explicit knowledge, there is a problem of process purity in the measurement of recollection and familiarity. Whereas remember was thought to be associated with recollection and know with familiarity, the SDT model specifies that both remember and know decisions are made on the basis of both recollection *and* familiarity.

Arndt and Reder (2002) provided evidence for dual processes by eliminating the FA portion of the mirror effect whilst leaving the HR portion intact. They did so by using lures that were plural forms of studied words in a word-frequency mirror effect experiment. Participants were informed that either the plural or singular form of a word would have

appeared on the study list. Arndt and Reder reasoned that participants would then be able to use their recollection to both reject and accept stimuli – if they remembered seeing a plural form at study when presented with a singular form, they could be sure the singular form was not on the study list. For this set of lures, the FA portion of the mirror effect did not occur. Arndt and Reder concluded that a two criterion model best explained recognition performance. A high criterion is placed on the strength-of-evidence scale above which an item is rated as old. A second, lower criterion is placed below which an item is identified as new. In-between these criteria familiarity is used to make identifications. This is not too dissimilar to Scott and Dienes (2008) familiarity model where above and below a certain level of familiarity, participants can explain that they used familiarity but in-between participants do not know the source of their judgements.

Other parallels exist between mirror-effect research and implicit-learning research. For instance, similarity has been implicated as a factor in each area. Ozubko and Joordens (2008) demonstrated that by increasing the similarity between classes of stimuli the FA part of the mirror effect can be reversed so that the FAR goes up from the non-distinct to the distinct stimuli category. The susceptibility of recollection to interference when familiarity is left untouched also parallels manipulations seen in implicit learning. Joordens and Hockley (2000) found that requiring speeded responding suppressed the HR portion of the mirror effect whilst leaving the FA portion untouched, suggesting that speeded responding affected recollection but not familiarity.

1.7 Unifying Implicit Learning and Recognition Memory

Despite the fact that the methods and frameworks are similar in both mirror-effect research and implicit learning research, there have been few attempts to integrate these areas. Combining recognition memory and implicit learning research would allow each to add to their armoury an extended set of theories and methods. To this end, questions such as "Is there a mirror effect in classification tasks?" and "How does implicit learning interact with the mirror effect?" are important to answer. If both classification performance and recognition performance share an underlying process such as familiarity, then theories of learning and memory must explain results found in both areas of research.

The experiments in the following chapters will take a step towards the goal of unifying the implicit learning and recognition memory by applying methods found in the study of the mirror effect to classification performance after implicit learning. Initially, a set of stimuli will be developed to be used in an implicit-learning paradigm (Chapter 2). This will be

further modified to draw on strength-based mirror-effect research and SDT in order to investigate the processes underlying implicit learning and recognition memory (Chapter 3). Alternative explanations for Chapter 3's findings will be considered (Chapter 4). Specific limitations of the findings will be addressed and the data considered in terms of specific memory models, including a MINERVA simulation (Chapter 5). Finally, the implications of the findings for implicit learning and recognition memory will be discussed (Chapter 6).

2 Chapter 2: Testing the Stimuli

2.1 Introduction

Most implicit-learning experiments use either the AG paradigm (e.g. Dienes & Scott, 2005; A. S. Reber, 1989) or the SRT paradigm (e.g. Norman et al., 2007; Rowland & Shanks, 2006; Song et al., 2008) However, these findings do not generalise to all cases of implicit learning. Specifically, neither paradigm offers a complete account of rule learning with natural words.

Unlike natural words, AG strings are not pronounceable and do not have underlying meaning. Higham (1997a) demonstrated that when test items were pronounceable participants were more sensitive to similarities between test items and study items than when the test items were not pronounceable. Learning of surface characteristics such as letter pairs could not account for this result, because each set of stimuli contained the same number of letter pairs. Participants are also sensitive to rule sets that are defined in terms of underlying characteristics of words such as frequency of occurrence in language and whether the words are nouns or verbs (Higham & Brooks, 1997). Natural words are more complex stimuli than AG strings because words and even the characteristics of words have deeper underlying meanings.

SRT stimuli are also less complex than natural words. SRT sequences are usually discussed with reference to their surface or statistical properties (e.g. J. Reed & Johnson, 1994). An SOC sequence, for example, is a sequence where the location of the third stimulus in a chain of three can be defined entirely by the location of the two preceding stimuli. Sometimes the relationship is probabilistic where the location of the third stimuli can be one of several different locations, but the set of possible locations is constrained by the preceding two stimuli. The letters in a natural word also follow similar rules, but learning words is more than just learning the possible sequences of letters. With words the sequence of letters represents an intrinsic meaning and set of characteristics, something not present in a SRT sequence.

As SRT and AG stimuli lack intrinsic meaning they cannot be used to embody some classes of rules. Consider the concepts of abstractness and concreteness, for example. Although words can be classified into one or other of those categories, the categories themselves have meaning beyond the surface characteristics of the words and this meaning is required in order to perform any kind of classification involving concreteness and

abstractness. Experiment 1 will therefore be an exploratory experiment to test the use of natural words in an implicit-learning task. The rule set to be learned by participants will be instantiated partly using the deeper characteristics of the words. Participants will be unable to perform the classification task using just the surface characteristics of the stimuli. In this way, the limitations of AG and SRT stimuli will be circumvented and more generalizable results can be obtained.

The second goal of Experiment 1 concerns the debate about measures of awareness. In implicit learning experiments, participants usually learn the rule set without being instructed to and do not develop explicit knowledge of what they have learned. Many subjective measures of awareness focus on either above-chance performance when participants claim to be guessing (guessing criterion) or relationships between confidence and accuracy (zero-correlation criterion) (Dienes, Altmann, Kwan, & Goode, 1995). There are different methods to calculate both the guessing criterion and the zero-correlation criterion. There is still much controversy about what these methods actually measure and around the sensitivity of different specific measures to conscious and unconscious factors. For example, Tunney and Shanks (2003) demonstrated that measures employing a binary choice were more sensitive to conscious knowledge than were those employing continuous rating scales. On the other hand, Dienes (2008a) found no difference in the sensitivity of different scales. More recently, there has been much debate about the use of wagering in order to discover whether people have explicit knowledge of their decisions (Dienes & Seth, 2010; Overgaard, Timmermans, Sandberg, & Cleeremans, 2010; Persaud, McLeod, & Cowey, 2007, 2008; Seth, 2008). The uncertainty about the sensitivity of different measures makes it difficult to pick an appropriate measure of awareness. Thus Experiment 1 compared a variety of methods of judging the awareness of participants in order to assess which measures are conservative and which are liberal in classifying participants as unaware.

2.1.1 Basis for the experiments.

The main experimental influence is drawn from an experiment by Higham and Brooks (1997). Natural words were included on a study list using a rule set. One rule denoted the frequency of the word and the other rule denoted the category of the word – either noun or verb. Words were included on the list by conjoining these two rules. Study List 1 included low-frequency nouns and high-frequency verbs whereas Study List 2 included high-frequency nouns and low-frequency verbs. Higham and Brooks reasoned that this rule set

would be hard to verbalise. This design also has the advantage that knowledge of only one rule of the set is orthogonal to the set as a whole. Knowledge of the conjunctive rule is needed for above-chance performance on the classification task. Additionally, the test items can be counterbalanced by rule set such that attributes of the specific stimuli can be controlled; that is, the rule-consistent words for one rule set act as rule-inconsistent words for the other rule set, thereby counterbalancing for surface characteristics such as word length.

In the study phase of Higham and Brooks' (1997) second experiment, participants rated words for understanding in order to ensure that each word was deeply encoded. This allowed the maximum opportunity for the participant to learn about the characteristics of each word. At test, participants completed a classification task or a recognition task. In the classification task participants made judgements based on the rule consistency of the words, whereas the recognition task required participants to identify words they had seen in the study phase. Test stimuli were a combination of words that had been in the study phase (old), new words that were rule consistent (i.e. NC words) and new words that were not rules consistent (i.e. NI words). In the classification task, participants performed above chance on both old and NC words, indicating that participants could identify rule-consistent stimuli. In the recognition task, participants misclassified NC words as old more than NI words, indicating that participants could not control the use of this knowledge. In a post-experimental questionnaire designed to determine whether the basis for performance was tacit, not a single participant demonstrated any knowledge of the rule set. Apart from noting that these results were consistent with a dual-process approach, the identity of these processes was not fully investigated.

Following the general design of Higham and Brooks (1997), Experiments 1 and 2 used a conjunctive rule-set. The study list in one condition contained common-concrete words such as *hotel* and rare-abstract words such as *tidal*. The second study list contained rare-concrete words such as *kite* and common-abstract words such as *written*. The complexity of this rule set is such that participants are unlikely to learn it explicitly, but as in Higham and Brooks (1997) participants should be able to identify rule-consistent stimuli.

Several task elements were introduced in Experiment 1 in order to assess participants' awareness of the knowledge they were using to discriminate the words. Dienes and Scott (2005) theorised that two types of knowledge are used in implicit learning experiments. Judgement knowledge is the knowledge of whether a given stimulus is rule consistent. Structural knowledge is knowledge about which features of a stimulus make it rule consistent. It is important to note that each type of knowledge can be conscious or

unconscious. For instance, a person may know that something is the case (explicit judgement knowledge) but not know *why* it is the case (implicit structural knowledge). Measures based on confidence ratings tend to measure judgement knowledge. Thus Experiment 1 included a confidence rating which could be used to calculate a variety of measures of judgement knowledge. The specific measures used were type-2 signal detection *d'* (see section 1.3.2 for definition); direct comparison of type-2 HRs and FARs (also see section 1.3.2); Goodman-Kruskal gamma (Nelson, 1984); Phi (Nelson, 1984); Pearson correlation and the Chan difference score (Dienes et al., 1995).

Measuring structural knowledge is more difficult. Dienes and Scott (2005) recommended a method in which participants state the basis of each of their choices. They gave participants a choice of four knowledge attributions - random chance, intuition, memory and rules. Random chance and intuition were theorised to represent mostly unconscious structural knowledge whilst memory and rules represented cases where the structural knowledge was mostly conscious. Note that each category can indicate a mix of conscious and unconscious knowledge, but the choice indicates the state of the *majority* of the knowledge used for that trial. Even with a memory attribution it is possible that some of the structural knowledge being used is implicit. Following this recommendation, Experiments 1 and 2 gave participants a choice of four attributions to communicate the basis of their decisions. Additionally, a post-experiment questionnaire was used to directly ask participants if they had any knowledge of the rule set.

2.2 Experiment 1

2.2.1 Predictions.

As this study is exploratory no predictions were made other than to state that participants will perform above chance when asked to discriminate NC words from NI words.

2.2.2 Method.

2.2.2.1 Participants.

Thirty-three participants from the University of Southampton (UoS) took part in the experiment for course credit or £5.

2.2.2.2 Materials and design.

The experiment consisted of a study phase followed by two test phases. For the study phase two study lists of 80 words each were used. The words were drawn from the MRC psycholinguistic database - see Wilson (1988). Words were classified as either common (frequency of 80+ per million) or rare (frequency of 1 or less per million) by Kucera-Francis written-frequency norms (Kucera & Francis, 1967). Each word was also classified as concrete (rating of 520 or more) or abstract by the MRC concreteness rating which merges data from several sources (Coltheart, 1981). Due to a shortage of words with a low concreteness rating in the database, abstract words were identified by the experimenter from the set of unrated words in the database. The words in this pool were split into four further word pools by combining their concreteness with their frequency. This resulted in four categories of words: common-concrete (CC) words (e.g. *hotel*), rare-abstract (RA) words (e.g. *tidal*), rare-concrete (RC) words (e.g. *kite*) and common-abstract (CA) words (e.g. *written*). These four word pools were screened and any words that could easily be interpreted as both concrete and abstract were eliminated. Two study lists were created by randomly selecting 40 words from each of the four word pools. Study List A consisted of 40 CC words and 40 RA words and Study List B consisted of 40 RC words and 40 CA words. Words were therefore assigned to each study list based upon a rule set which combined two factors to make rule-consistent words. Each word on a study list could be rule consistent in one of two ways (i.e. common-concrete or rare-abstract on Study List A and rare-concrete or common-abstract on Study List B). In the study phase, participants were shown one of these two lists and asked to rate each word for understanding. Half of the participants were given Study List A and the other half were given Study List B. Thus rule set was counterbalanced across participants.

Words used in the test phases were all new words not seen in the study phase. Words either complied with the study-phase rule-set (NC words) or did not comply with the rule set (NI words). Two test lists were created – Pair List A and Pair List B. From each of the four word pools 80 words were drawn for a total of 320 words. From the set of 320 words one set of 160 word pairs (e.g. hotel/kite) were created by pairing words from different categories together. Not all categories were paired together - there were 40 CC/RC pairs, 40 CC/CA pairs, 40 RA/CA pairs and 40 RA/RC pairs. There are two important things to note about this set of pairings. Firstly, no matter which study list a participant had seen one word in the pair was always a NC word and the other word in the pair was always a NI word. This was

the case because no CC words were paired with RA words and no RC words were paired with CA words. Secondly, each type of rule violation (frequency or concreteness) was equally represented because the words in each word pair were always matched on either frequency or concreteness but not both (e.g. if both of the words in a pair were rare, then one word was abstract and the other was concrete). The two test lists were compiled using 20 word pairs of each type, making 80 word pairs on each list. In the first test phase, participants were given the word pairs one at a time and asked to complete a classification task in which they had to choose the NC word in each pair. They were then asked to make a confidence judgement about their classification. In the second test phase, participants were asked to complete the classification task, give confidence ratings and also to provide a judgement about the basis of their decision. Everyone participated in both test phases. Assignment of test list to test phase was counterbalanced across all participants. Once created each list was not changed, although the presentation order of the word pairs was randomised anew for each participant.

A questionnaire was also used to assess verbalisable knowledge of the rule set. The questionnaire consisted of five questions, with early questions being vague (e.g. What did you think the rules were?) and later questions getting increasingly more direct (e.g. participants were told that the rule was conjunctive and given a list of possible rules from which they were asked to pick two). See Appendix B for the questions used (Question 1b was not used in this experiment).

2.2.2.3 Procedure.

Ethics approval for all Experiments was sought and granted from Southampton University's School of Psychology Ethics Committee. All participants signed consent forms before completing the experiment on an Apple Macintosh computer using the Revolution program[2]. The study phase consisted of 80 words from one of the two study lists. Words were displayed in sets of eight. Presentation order of the words was randomised separately for each participant. Attention was assured by the participant rating each word for how well they understood its meaning – a rating of one indicated that they did not understand the meaning at all whilst a rating of four indicated that they fully understood the meaning of the word. Participants were not told about the rule set at this stage.

After the study phase, participants were informed of the existence of a rule set but not what it was. The test phases were then administered. In each test phase, word pairs from one

[2] Revolution is a program in which run applications, and can be purchased at www.runrev.com

of the test lists were presented on the screen in a random order. One word from the pair was randomly picked to be displayed on the left of the screen whilst the other word was displayed on the right. In Test Phase 1, participants were asked to complete a classification task in which they had to indicate which word of the pair was the rule-consistent word. After making this judgement, participants had to provide a confidence rating about the correctness of their answer on a scale of 50-100 by typing it in a response box. For Test Phase 2, participants performed both the classification and confidence judgements and in addition provided a judgement about the basis of their decision from a choice of random chance[3], intuition, memory and rules using a radio button. The test was self-paced - all judgements were present on screen simultaneously and participants initiated the next trial by clicking a button with the mouse. Participants were prevented from moving on until all information had been entered. Note that the test lists were counterbalanced, but the tasks were not – the attribution task always appeared in the second block.

Once both test phases had been completed, the questionnaire (Appendix B) was administered by computer to probe for understanding of the rule set. Questions were administered one at a time.

2.2.3 Results.

Identical results were obtained using Study List A and Study List B therefore all analyses were collapsed across study list. Note that here and in all *t* tests reported in this document *d* refers to the measure of effect size Cohen's *d*, not to be confused with the signal-detection measure *d'* which will appear in later chapters.

2.2.3.1 Accuracy and confidence ratings.

Mean accuracy is presented in Table 2.1. Classification of NC words over a chance level of 50% would signify that participants had learned the rule set. Accuracy was compared to 50% using the lower bound of the 95% confidence interval, as estimated using 1.96 standard errors. As can be seen from Table 2.1, accuracy was above chance in Test Phase 1, Test Phase 2 and overall.

[3] The participants were given the attribution of random chance in order to avoid any associations and beliefs a participant might have about guessing. However the intention is that this attribution represented guessing and it will be used to calculate the guessing criterion. Thus for the remainder of the document the random chance attribution will be referred to as the guess or guessing attribution

Table 2.1

Accuracy and Confidence in Experiment 1 (SE in brackets)

Test Phase	Mean Accuracy (%)	Mean Confidence
Phase One	58.98 (1.06)*	60.08 (1.41)
Phase Two	55.64 (1.28)*	59.65 (1.30)
Overall	57.31 (0.88)*	59.86 (1.33)

* = Lower bound of 95% confidence interval above chance level of 50%.

2.2.3.2 Attribution performance.

See Table 2.2 for accuracy and confidence by attribution choice as well as proportion of usage of each attribution. Accuracy for each attribution type was the proportion of correct responses conditionalised on the number of responses for that attribution.

Table 2.2

Mean Accuracy, Confidence and Proportion of Use for Attribution Choices in Experiment 1 (SE in brackets)

Attribution	Accuracy	Confidence	Proportion
Guessing	53.98 (2.12)	52.42 (0.79)	.46 (.04)*
Intuition	58.89 (2.78)	61.62 (1.24)	.30 (.03)
Memory	62.40 (5.43)	72.97 (2.93)	.09 (.02)*
Rules	58.29 (4.51)	71.37 (2.39)	.15 (.03)*

* = Proportion different from .25 by plus/minus 1.96 standard errors.

Because the sum of the proportion variable necessarily equals one, comparing individual proportions directly in the same analysis is problematic. Therefore, the proportions of attributions were separately compared to the chance-level of attribution use of .25 using 1.96 standard errors (see Table 2.2). Only intuition was no different from .25. Participants appeared to have used guess more often than would be expected and memory and rules less often than would be expected on the hypothesis that they are all used equally.

Accuracies for the attribution types were entered into a one-way, repeated-measure analysis of variance (ANOVA). Only those participants who had used all four attributions were entered into the ANOVA. Dienes and Scott (2005) found higher accuracy for memory

and rules attributions than for guess and intuition attributions. Here there were no differences in accuracy by attribution at all, $F(3, 51) = 1.11$, $p = .35$. Confidence by attribution was also entered into a one-way repeated-measure ANOVA. There was a main effect of attribution, $F(3, 51) = 26.22$, $p < .001$, $\eta^2 = .61$. Pairwise comparisons further clarified the main effect. Guess confidence ratings were smaller than intuition, memory and rules confidence ratings, $F(1, 17) = 54.03$, $p < .001$, $\eta^2 = .76$, $F(1, 17) = 32.79$, $p < .001$, $\eta^2 = .70$ and $F(1, 17) = 44.56$, $p < .001$, $\eta^2 = .72$ respectively. Intuition confidence ratings were smaller than memory and rules confidence ratings, $F(1, 17) = 17.38$, $p = .001$, $\eta^2 = .51$ and $F(1, 17) = 23.87$, $p < .001$, $\eta^2 = .58$. Memory and rules confidence ratings were not different, $F < 1$.

2.2.3.3 Measures of awareness.

2.2.3.3.1 Questionnaire responses.

Similar to Higham and Brooks (1997) the questionnaire indicated that no participant developed knowledge of the rule set, even when directly told that the rule was a conjunctive rule and given a list of possible factors.

2.2.3.3.2 Guessing criterion.

If accuracy performance was above chance in conditions where participants claimed to be guessing, then it can be said that they have implicit knowledge of the rule set. Two versions of the guessing criterion can be tested – accuracy when participants give a 50% confidence rating and accuracy when participants give guess attributions. See Table 2.3 for accuracy means. One participant only gave one 50% confidence rating and was excluded from the analysis. When participants gave 50% ratings accuracy was above chance. Guess attribution accuracy was not above chance (see Table 2.3).

Table 2.3

Percentage Accuracy for 50% Confidence Ratings and Guess Responses in Experiment 1 (SE in brackets)

Phase Number	50% confidence	Guess
Phase One	55.73 (2.03)*	-
Phase Two	55.87 (2.47)*	53.98 (2.12)

* = Lower bound of 95% confidence interval above chance level of 50%.

2.2.3.3.3 Zero-correlation criterion.

The zero-correlation criterion was measured in two ways. The first was to calculate a Pearson correlation coefficient between accuracy and confidence for each test phase. A positive correlation indicated explicit knowledge. The second method was the Chan difference score (Dienes et al., 1995) which involves comparing confidence of correct answers to confidence of incorrect answers. If confidence for correct answers is higher than confidence for incorrect answers, then knowledge is said to be explicit.

For both Test Phase 1 and Test Phase 2 there was no correlation between confidence and accuracy, $r(33) = .03$, $p = .85$, $r^2 = 0.00$ and $r(33) = -.09$, $p = .61$, $r^2 = .01$ respectively. This indicated unconscious knowledge by the zero-correlation criterion. Paired t tests measuring the Chan difference score indicated that correct answer confidences were higher than incorrect answer confidences in both Phase 1, $t(32) = 3.34$, $p = .002$, $d = 0.22$ and Phase 2, $t(32) = 2.19$, $p = .04$, $d = 0.09$. See Table 2.4 for correct and incorrect confidence means.

Table 2.4

Correct and Incorrect Confidence Means in Experiment 1 (SE in brackets)

Test Phase	Correct	Incorrect
Phase One	60.69 (1.49)	59.08 (1.33)
Phase Two	59.90 (1.31)	59.22 (1.31)

2.2.3.3.4 Other measures.

In this section, other measures are reported to judge the awareness of the participants. The first measure is a type-2 receiver operating characteristic (ROC) curve. A type-2 ROC curve plots a set of type-2 HRs and FARs (see section 1.3.2) against each other. The set is created by setting the split point for high and low confidence at different levels. If the curve bows away from the 45 degree diagonal line then this suggests that participants were aware of when they were correct and when they were incorrect. The criteria used were 50, 60, 70, 80, 90 and 100. As this gives us a small number of points the non-parametric sign test was used to check if the ROC deviated from the chance line (see Figure 2.1 and Figure 2.2 for ROC curves). The type-2 HR was greater than the type-2 FAR for Phase 1, sign test (N = 6), $p = .04$, but not Phase 2, sign test (N = 6), $p = .07$. This indicated that participants had explicit knowledge in Phase 1 but not in Phase 2.

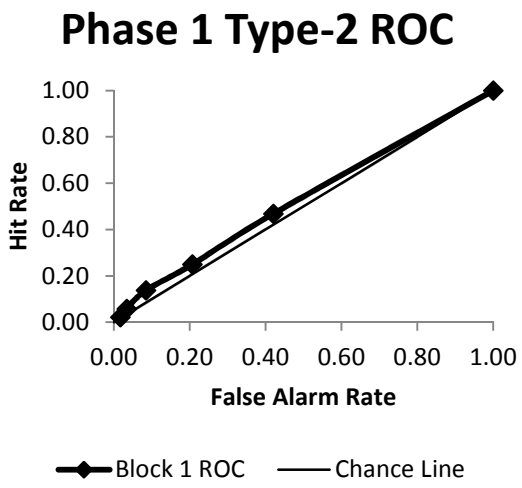

Figure 2.1. Type-2 ROC for Phase 1, Experiment 1.

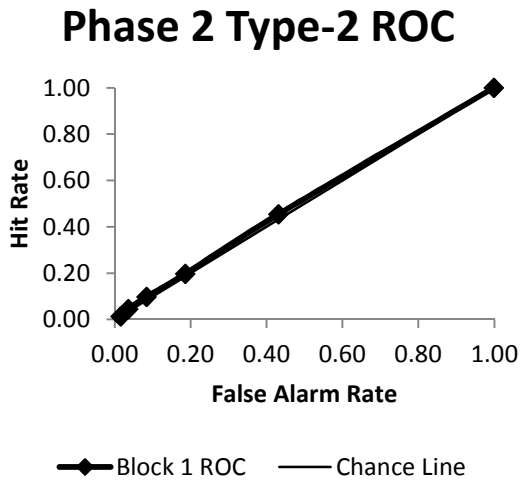

Figure 2.2. Type-2 ROC for Phase 2, Experiment 1.

Phi and type-2 *d'* both require classifying what is high confidence and what is low confidence for each participant. This was achieved by splitting confidence for each participant around the median for that participant. Type-2 *d'* was calculated by subtracting the z type-2 FAR from the z type-2 HR. Phi is a correlation computed on a 2 x 2 table (see Nelson, 1984). A Phi of greater than 0 would indicate that confidence and accuracy are related. Goodman-Kruskal gamma compares the product of the number of concordances and the number of discordances at each level of a criterion – the number of high confidences to correct answers and low confidences to incorrect answers are concordances whilst the number of low confidences to correct answers and high confidences to incorrect answers are discordances. Any gamma higher than zero reflects a relationship between accuracy and

confidence. All three measures were compared with zero using the lower bound of the 95% confidence interval (Table 2.5). Type-2 *d'* and Phi both indicated explicit knowledge in Phase 1 and implicit knowledge in Phase 2, whilst gamma indicated explicit knowledge in both phases.

Table 2.5
Values for Awareness Measures in Experiment 1(SE in brackets)

Test Phase	*d'*	Phi	Gamma
Phase One	0.24 (.07)*	0.09 (.02)*	0.15 (.04)*
Phase Two	0.08 (.06)	0.02 (.02)	0.09 (.04)*

* = Lower bound of 95% confidence interval above chance level of 0.

2.2.4 Discussion.

Consistent with Higham and Brooks (1997) participants could identify rule-consistent words at above-chance levels. This would not be possible unless participants were in some way sensitive to the conjunction of concreteness and frequency that composed the rule set. It is unlikely that participants would have spontaneously looked for rules linking the words as nothing in the study task encouraged them to do so. Thus whatever creates this sensitivity must have occurred incidentally. Broadly speaking, there are two possible explanations for such an incidental process. An automatic learning mechanism could become sensitive to the underlying rule set as A.S. Reber (1967) initially claimed. Alternatively, test items could be compared to memory for some or all of the study items, such as with a chorus-of-instances approach (e.g. Vokey & Higham, 2005). Such a process need not involve explicit memory of any kind (see Higham & Vokey, 1994). The results of this experiment are equivocal as to which of these could be the case. Further discussion of this point will be reserved for later chapters.

Participants appeared to use attributions appropriately. Guess attributions were the most common choice whilst memory and rules were used infrequently. This suggests that participants were not primarily relying on explicit strategies such as memory for parts of the words or knowledge of the rule set to make their decisions. However, accuracy was above chance in both memory and rules attributions and this accuracy was associated with a high level of confidence. Although it is possible that participants picked memory and rules inappropriately, studies by Higham, Bruno and Perfect (2010) as well as Dienes and Scott (2005) suggest that attributions do reflect differences in participants' cognitive states.

Memory and rules attributions need not reflect explicit structural knowledge. Theories from recognition memory such as MINERVA (Hintzman, 1984, 1986) can produce a strong feeling of familiarity for a stimulus that is created through the amalgamation of many instances in memory. This could lead to a memory attribution without the ability to verbalise the specific details of the memory. Rules attributions could reflect the usage of a rule set other than that presented in the study phase. This could lead to above-chance performance with rules attributions in one of two ways. Participants could be using a rule set that is related to the actual rule set such as chunk strength, resulting in rules attributions. Alternatively, the fact that participants have been told that a rule set is present may lead them to choose the rule attribution for high confidence responses, reasoning that they must have some knowledge of the rule set even if they cannot verbalise that knowledge. However, as reflected in the questionnaire no participant had verbalisable knowledge of the rule set. Given the low usage of memory and rules attributions, the use of attributions associated with explicit structural knowledge was examined in more detail in Experiment 2.

Table 2.6 summarises the results for the subjective measures of awareness. The guessing criterion as measured with the guess category indicated chance performance, whilst the guessing criterion with the 50% confidence ratings indicated above-chance performance. Thus different applications of the guessing criterion yielded different conclusions about implicit knowledge. This could be due to a sensitivity difference between the two scales (Tunney & Shanks, 2003) although it may be that the percentage scale is the more sensitive of the two (Dienes, 2008a).

Table 2.6
Awareness Measures Summary for Experiment 1

Test Phase	Guess	50% Confidence	Pearson	Chan	Sign test/ROC	*d'*	Gamma	Phi
Phase One	-	I	I	E	E	E	E	E
Phase Two	No	I	I	E	I	I	E	I

Note. I = Test indicated implicit knowledge E = Test indicated explicit knowledge No = Test indicated no knowledge.

For the zero-correlation criterion, only the Pearson correlation consistently identified no relation between accuracy and performance. The Chan, sign test and Gamma measures all

identified a relation between accuracy and performance, whilst the *d'* and Phi measures identified a relation in Phase 1 but not Phase 2. The choice of measure of awareness is thus an important factor in experiments. From this set of results, it would appear that the guessing criterion and Pearson correlation are more liberal in declaring implicit knowledge whilst the others are perhaps more conservative. Which measure is best to use in any one situation depends on what stance the experimenter takes on the "Consciousness as King" (J. Reed & Johnson, 1998; Shanks & St. John, 1994) issue – if the stance is that explicit knowledge is assumed to exist but implicit knowledge is not, then conservative measures should be used. If implicit knowledge is assumed to exist and operate, then liberal measures are a reasonable choice. It is worth remembering that all of these measures only track participants' confidence in their answers, so explicit knowledge in this case refers to explicit judgement knowledge. As Dienes and Scott (2005) point out, it is possible for a person to be confident in their response, but unable to explain the reasons behind their confidence. In Experiment 1 this would seem to be the case, as no participant could explain the rule set.

In conclusion, the rule set behind these stimuli appeared to be learned without any deliberate effort to do so making the stimuli suitable for use in implicit-learning experiments.

2.3 Experiment 2

2.3.1 Introduction.

Distraction is a commonly used manipulation in implicit learning and recognition memory studies. Jacoby (1991) used a distraction task to reduce recollection in order to measure the effects of familiarity. Some studies have found that distraction at test reduces explicit-knowledge-based performance but not implicit-knowledge-based performance (Dienes & Scott, 2005; Foerde et al., 2007; Frensch, Lin, & Buchner, 1998). Implicit learning appears to be resilient to distraction in the study phase (Frensch et al., 1998). In both recognition memory and implicit learning literature, distraction seems to suppress the influence of recollection and explicit knowledge whilst leaving familiarity and implicit knowledge intact. Thus introducing a distraction manipulation here served two purposes. It ensured participants' ability to identify rule-consistent stimuli in Experiment 1 was primarily due to implicit knowledge or at least non-verbalisable knowledge of some type. The distraction manipulation also served as a check for what kind of knowledge is represented by memory and rules attribution – explicit knowledge or knowledge based on recollection should be reduced by distraction whilst implicit knowledge should remain intact.

As discussed above, theories and models from recognition memory and implicit learning are often similar – for instance the chorus-of-instances (Higham & Vokey, 1994; Vokey & Brooks, 1992) approach is very similar to the global-memory model MINERVA (Hintzman, 1984, 1986). MINERVA has even been used to show that a dual-process explanation is not needed to explain some dissociations from implicit-learning experiments (Jamieson et al., 2010). Performance in implicit-learning experiments requires the use of memory in some way as either the rule set or some aspect of the stimuli must be remembered to make a successful classification judgement. It has been suggested in the implicit-learning literature that familiarity is the mechanism which enables successful classification judgements (Scott & Dienes, 2008). Memory literature, on the other hand, often cites familiarity as a component of recognition judgements (Jacoby, 1991; Yonelinas, 2002). Both literatures seem to suggest that familiarity is an important mechanism. The difference is that recognition literature employs familiarity to identify words seen in the study phase, whilst implicit-learning literature employs familiarity to identify rule-consistent words. With this in mind, Experiment 2 included words from the study list ("old" words) at test. This inclusion makes Experiment 2 more similar to recognition experiments, allowing both recognition memory and implicit learning to be examined side-by-side in the same experiment.

2.3.2 Predictions.

Frensch, Lin et al. (1998) found that implicit performance was not affected by distraction at study or test. Explicit performance *was* affected by distraction at test. As the learning is primarily implicit in this experiment it was predicted that distraction at study will not affect performance for NC words. Distraction at test would affect performance for old words, but should leave performance for NC words intact. In the attribution categories of memory and rules, performance for NC words would not be affected if such attributions were made on the basis of implicit knowledge but should be reduced if the attributions were made on the basis of explicit knowledge.

2.3.3 Method.

2.3.3.1 Participants.

Thirty-two participants took part in the experiment recruited from the UoS for course credit or £5 cash.

2.3.3.2 Materials.

The same study lists were used as in Experiment 1. The same test lists were also used with some modifications. In order to include words seen in the study phase at test ("old" words) the test lists had to be associated with a study list. Thus the same two test lists used in Experiment 1 became four test lists for Experiment 2 – two test lists associated with Study List A and two test lists associated with Study List B. The 80 old words were separately paired with 80 NI words drawn from the relevant word pools. As for the previously existing word pairs, the additional NI words had each type of rule violation equally represented (i.e. the NI word matched the old word on only one of concreteness or frequency). All permissible types of pairing were equally represented so long as an old word was paired with an NI word. For Study List A this resulted in 20 CC/RC pairs, 20 CC/CA pairs, 20 RA/RC pairs and 20 RA/CA pairs. For Study List B this resulted in 20 RC/CC pairs, 20 RC/RA pairs, 20 CA/CC pairs and 20 CA/RA pairs. For each study list 40 word pairs were then added to the first test list and 40 to the second test list, with equal numbers of each type of word pair being used on each test list (i.e. 10 of each type were added to each test list). This resulted in the original two test lists of 80 word pairs becoming four test lists of 120 word pairs.

2.3.3.3 Design.

A 2 x 2 design was used. Distraction at study was manipulated between-subjects such that half the participants were distracted at study whilst the other half were not. Distraction at test was manipulated within-subjects such that all participants were given a test phase in which they were distracted and a test phase in which they were not distracted. Test distraction order was counterbalanced such that half the participants had the distraction phase first whilst the other half had the distraction phase second. Thus there were four conditions – distracted at study/distracted at test; distracted at study/not distracted at test; not distracted at study/distracted at test; and not distracted at study/not distracted at test. Assignment of test list to test phase was counterbalanced across participants. Study list was also counterbalanced across participants.

2.3.3.4 Procedure.

The whole task was administered on an Apple Macintosh computer. The study phase was the same as Experiment 1 except for two changes. In Experiment 1, participants may

have explicitly noticed commonalities between the words because they were presented in sets of eight. To ensure this did not happen here, words were displayed one at a time. Additionally, different participants in Experiment 1 may have spent varying amounts of time viewing the words, or the same participant may have spent varying amounts of time on each set of eight words. Thus, in this experiment, a display limit was introduced for each word. Words were displayed one at a time for 1.5 seconds each, with meaning ratings being requested after each word. Half of the participants were distracted in the study phase by a number counting task (Craik, 1982; Jacoby, 1991). During the study task, a series of numbers were played through a set of headphones at a rate of one every 1.5 seconds. The participants were asked to press the space bar every time they heard three odd numbers in a row. The first time and then every fourth time that participants failed to respond at the appropriate time a box appeared on the screen reminding them to attend to the audio task.

The test phase consisted of two blocks of 120 words pairs each. The instructions and task were the same as the second block of Experiment 1 so participants picked rule-consistent words and provided confidence and phenomenological ratings. All participants were given a distraction task in one of the two test phases. The distraction was the same as was used in the study phase. This was counterbalanced such that half of the participants were distracted in Test Phase 1 whilst the other half were distracted in Test Phase 2. Note that this means only half the participants were distracted at study, whilst all of the participants were subjected to both distraction and no distraction at test. After the second test phase participants were presented with the same questionnaire as in Experiment 1.

2.3.4 Results.

There were no effects involving distraction at study or study list so analysis in this section is collapsed across these variables.

2.3.4.1 Accuracy performance.

Accuracy was calculated for old/NI and NC/NI pairs, where accuracy represented the extent to which participants correctly chose the old or NC word. Above-chance accuracy was tested for as in Experiment 1. There was above-chance accuracy for both old and NC stimuli regardless of whether the participant was distracted at test or not (see Table 2.7).

Table 2.7

Overall Accuracy and Confidence Means by Test Distraction and Pair Type in Experiment 2 (SE in brackets)

Test distraction and pair type	Accuracy	Confidence
Not distracted at test		
Old/NI	84.72 (2.31)*	80.34 (1.72)
NC/NI	57.07 (1.44)*	62.43 (1.82)
Distracted at test		
Old/NI	83.66 (1.99)*	79.22 (1.58)
NC/NI	56.81 (1.16)*	61.75 (1.84)

Note. Old = words seen at study, NC = New-consistent words NI = New-inconsistent words.

* = Lower bound of 95% confidence interval above chance level of 50%.

Accuracy was entered into a 2 x 2 repeated measures ANOVA with pair type (old/NI versus NC/NI) and test condition (distracted versus not distracted). There was a main effect of pair type, $F(1, 31) = 200.26$, $p < .001$, $\eta^2 = .86$ indicating that old/NI accuracy was higher than NC/NI accuracy. There were no other effects, highest $F < 1$.

Confidence was entered into a 2 x 2 ANOVA with pair type and test condition. There was only a main effect of pair type, $F(1, 31) = 160.82$, $p < .001$, $\eta^2 = .84$ indicating that confidence in old word choices was higher than confidence in NC word choices. There were no other significant effects, highest $F(1, 31) = 1.42$, $p = .24$.

2.3.4.2 Attribution performance.

In order to analyse the effects of distraction on each component of performance, correct and incorrect answers in each attribution were analysed. The answers were analysed as a proportion of all total answers. See Table 2.8.

Table 2.8

Attribution Contribution to Overall Performance in Experiment 2 (SE in brackets)

Answer, distraction and pair types	Guess	Intuition	Memory	Rules
Correct answers				
Not distracted at test				
Old/NI	0.09 (.01)	0.11 (.01)	0.58 (.04)	0.07 (.02)
NC/NI	0.22 (.02)	0.19 (.02)	0.09 (.02)	0.07 (.01)
Distracted at test				
Old/NI	0.11 (.02)	0.13 (.01)	0.52 (.04)	0.08 (.03)
NC/NI	0.24 (.02)	0.17 (.02)	0.09 (.02)	0.07 (.01)
Incorrect answers				
Not distracted at test				
Old/NI	.06 (.01)	.04 (.01)	.04 (.01)	.02 (.01)
NC/NI	.20 (.02)	.12 (.01)	.06 (.02)	.04 (.01)
Distracted at test				
Old/NI	.06 (.01)	.05 (.01)	.04 (.01)	.02 (.01)
NC/NI	.22 (.02)	.13 (.01)	.05 (.01)	.03 (.01)

Note. Old = words seen at study, NC = New-consistent words NI = New-inconsistent words.

Many participants did not use the rules category so it was not analysed. The other attributions were entered into separate 2 x 2 ANOVAs with test condition (distracted versus not distracted) and pair type (old/NI versus NC/NI). Correct and incorrect answers were analysed separately.

The correct guess attribution analysis revealed only a main effect of pair type $F(1, 31) = 77.40$, $p < .001$, $\eta^2 = .71$, indicating that correct guess attributions contributed more to accuracy for NC/NI ($M = .23$, $SE = .02$) pairs than for old/NI pairs ($M = .10$, $SE = .01$). There were no other effects, highest $F(1, 31) = 2.76$, $p = .11$. Similarly, the correct intuition analysis also revealed only a main effect of pair type $F(1, 31) = 14.32$, $p < .001$, $\eta^2 = .32$ reflecting a greater contribution from intuition attributions for NC/NI ($M = .18$. $SE = .01$) pairs than for old/NI pairs ($M = .12$, $SE = .01$). There were no other effects, highest $F(1, 31) = 3.53$, $p = .07$.

The correct memory attribution analysis revealed main effects of pair type $F(1, 31) = 121.26$, $p < .001$, $\eta^2 = .80$ and test condition $F(1, 31) = 7.06$, $p = .01$, $\eta^2 = .18$ and an

interaction between pair type and test condition $F(1, 31) = 5.95$, p = .02, $\eta^2 = .16$. The pair-type main effect indicated a greater contribution of memory attributions to accuracy for old/NI pairs ($M = 55$, $SE = .04$) than for NC/NI pairs ($M = .09$, $SE = .02$). The interaction indicated that distraction reduced the performance due to memory attributions, but only for old/NI pairs. The distraction main effect is likely to be an artefact of the interaction, as visual inspection of Table 2.8 would suggest that the only difference due to test condition is the change in the old/NI data.

For incorrect guess attributions, there was a main effect of pair type $F(1, 31) = 83.22$, $p < .001$, $\eta^2 = .73$. This reflected more incorrect guess attributions for NC/NI ($M = .21$, $SE = .02$) pairs than for old/NI pairs ($M = .06$, $SE = .01$). There were no other effects, highest $F < 1$. There were more incorrect intuition attributions for NC/NI pairs ($M = .13$, $SE = .01$) than for old/NI pairs ($M = .05$, $SE = .01$,) and more incorrect memory attributions for NC/NI pairs ($M = .06$, $SE = .01$) than for old/NI pairs ($M = .04$, $SE = .01$), $F(1, 31) = 49.60$, $p < .001$, $\eta^2 = .61$, $F(1, 31) = 9.76$, $p = .004$, $\eta^2 = .24$ respectively. There were no other effects highest, $F(1, 31) = 1.96$, $p = .17$

2.3.4.3 Awareness measures.

2.3.4.3.1 Questionnaire responses.

As in Experiment 1, the questionnaire indicated that no participant developed knowledge of the rule set.

2.3.4.3.2 Guessing criterion.

Using 50% confidence responses to test awareness against the guessing criterion, only NC/NI accuracy in the distracted test condition was above chance (see Table 2.9). Using guess responses to test awareness against the guessing criterion, only the old/NI not distracted at test accuracy was above chance.

Table 2.9

Guessing Criterion Data from Experiment 2 (SE in brackets)

Pair Type	50 Confidence	Guess
Not distracted at test		
Old/NI	56.52 (4.37)	58.57 (4.18)*
NC/NI	51.82 (2.01)	53.59 (2.03)
Distracted at test		
Old/NI	60.37 (5.33)	55.93 (5.44)
NC/NI	55.19 (2.33)*	52.42 (1.74)

Note. Old = words seen at study, NC = New-consistent words NI = New-inconsistent words.

* = Lower bound of 95% confidence interval above chance level of 50%.

2.3.4.3.3 Zero-correlation criterion.

The Chan difference score was used to measure the zero-correlation criterion. This was chosen as a measure that was simple and intuitive whilst incorporating information about correct and incorrect answers. In all conditions for both pair types, correct confidence was higher than incorrect confidence – old distracted $t(29) = 9.36$, $p < .001$, $d = 1.56$; old not distracted $t(30) = 9.66$, $p < .001$, $d = 1.55$; NC distracted $t(31) = 3.31$, $p = .002$, $d = 0.14$; NC not distracted $t(31) = 5.63$, $p < .001$, $d = 0.22$ – see Table 2.10 for means.

Table 2.10

Zero-correlation Data from Experiment 2 (SE in brackets)

Distraction and pair types	Correct Confidence	Incorrect Confidence
Not distracted at test		
Old/NI	83.00 (1.66)	61.71 (2.48)
NC/NI	63.42 (1.80)	61.04 (1.87)
Distracted at test		
Old/NI	81.50 (1.58)	62.70 (2.14)
NC/NI	62.50 (1.80)	60.94 (1.90)

Note. Old = words seen at study, NC = New-consistent words NI = New-inconsistent words.

2.3.5 Discussion.

Participants identified rule-consistent words at above-chance levels. Participants' performance for choosing old words over NI words in Experiment 2 was much better than that for choosing NC words over NI words. Both old and NC words benefit from the implicit effects of being rule-consistent. Old words could benefit from two additional influences. Participants could have an explicit recollection of seeing the old word on the study list. Alternatively, a feeling of having seen the word before without an explicit recollection of the word could be experienced (i.e. familiarity). Thus, identification of NC words is likely to be driven by primarily implicit processes and identification of old words is likely to be driven by a mixture of explicit and implicit processes. Higham and Brooks (1997) mentioned that a criticism of this type of study is that old words are included at test, which is not the case with many implicit learning studies. Experiments 1 and 2 provided a test of this criticism– the levels of performance for NC words are very similar in both experiments. Thus it appears that the presence of old words does not change the pattern of results.

That distraction at study had no effect on performance for NC words was to be expected - Frensch, Lin et al. (1998) found a similar result. However, the fact that distraction at study did not affect performance for old words was unexpected. Since participants gave self-paced meaning ratings, participants could pause if they had heard two odd numbers in a row. Once the next number had been heard participants could then rehearse the word they had just seen before giving a rating. This would render the distraction less effective. An alternative explanation is that the meaning ratings led to the study-list words being deeply encoded which may have protected against the effects of distraction.

Rules attributions were rarely made, reflecting the fact that no participant could explain the rule set. There were more memory ratings for old words than for NC words and there were more guess and intuition ratings for NC words than for old words. Distraction at test reduced only the contribution of performance in the memory category for old words. This is consistent with ideas from the recognition memory literature that distraction reduces conscious recollection of words but does not reduce the contribution of familiarity (Jacoby, 1991). In this case the contribution of familiarity to identification of old words is reflected by the guess and intuition contributions to old word accuracy. Distraction at test did not reduce the performance contribution of guess and intuition to accuracy for NC words, consistent with results suggesting that implicit knowledge is hard to disrupt (Dienes & Scott, 2005; Foerde et al., 2007; Frensch et al., 1998). This raises a question – is the same underlying

process driving the guess and intuition responses for the old word responses and also for the NC word responses? In both cases, there was no change in responding by distraction consistent with ideas of implicit learning and of familiarity. The evidence presented here does not allow a firm conclusion, so further discussion will be reserved for later chapters.

Each way of judging the guessing criterion identified chance performance in three out of four cases. This suggests that participants were using the guess and 50% confidence categories appropriately most of the time. On the other hand, the zero-correlation criterion indicated that participants were aware of the basis of their choices in all conditions. As no participant identified the rule set in the questionnaire and the rules category was rarely used, a positive result from the zero-correlation criterion suggests that they had conscious judgement knowledge (Dienes & Scott, 2005) rather than conscious structural knowledge. This is consistent with the idea that performance in classification tasks can be driven by a "feeling of rightness" which could be labelled familiarity (Scott & Dienes, 2008).

2.4 Conclusions from Experiments 1 and 2

AG and SRT experiments often cite knowledge of surface characteristics such as chunks as the basis of performance. Implicit learning involving natural words could be accomplished by a similar mechanism. However, natural words are richer stimuli than AG and SRT strings as they have a deeper meaning than just a sequence of letters. This makes it difficult to generalise from learning using AG or SRT stimuli to learning involving natural words. Using natural words as stimuli circumvents this limitation and enables the investigation of learning involving both surface features and conceptual features. The experiments presented here demonstrated that rule sets including such features could indeed be learned and employed.

The results using these stimuli are similar to results using other materials (e.g., AG strings) in implicit-learning literature. As the stimuli in Experiments 1 and 2 were natural words, further studies utilising them can make a direct comparison with results from those recognition-memory studies that also use natural words. There are already some studies that attempt to explain patterns of data found in implicit-learning experiments using models from recognition-memory literature, such as MINERVA (e.g. Jamieson & Mewhort, 2009a). There are also parallels between implicit learning theories such as the chorus-of-instance theory and recognition-memory models such as MINERVA and REM (Shiffrin & Steyvers, 1997). All of these theories use a comparison to all words seen on the study list in order to

explain performance in either recognition or classification tasks. This suggests a question – do the same processes drive performance in implicit learning and recognition memory? The next few chapters will address this question by using the natural word stimuli in both recognition and classification tasks. In particular two areas of interest will be investigated:

1) What exactly is learned in implicit learning? Can models of recognition memory such as MINERVA (Hintzman, 1984, 1986) and REM (Shiffrin & Steyvers, 1997) help explain what drives performance in classification tasks? This is especially relevant for explanations that hinge upon chorus-of-instance like explanations, because memory models such as MINERVA use this kind of mechanism.
2) In what areas can implicit learning and recognition literatures develop more cross-talk? For instance, if familiarity is the mechanism through which people make decisions in classification tasks, does this familiarity share the characteristics of familiarity as discussed in the recognition literature? If so, then cross-talk between recognition memory and implicit learning would be fruitful. If not, then it is important to be clear about which type of familiarity is being discussed and how implicit-learning familiarity differs from recognition familiarity.

3 Chapter 3: A Signal Detection Development of the Paradigm

3.1 Introduction

Chapter 2 tested the materials to ensure that participants learned the underlying rule set. This chapter refines the paradigm so that a signal-detection analysis can be used to separate the contributions of rule knowledge and memory to classification tasks and recognition-memory tasks. This separation will enable a closer study of the role of familiarity in each task.

3.1.1 Basis for the experiments.

In recognition literature, a mirror effect is said to occur when some change in a variable results in increased performance manifested by an increase in HRs and a decrease in FAs. Higham and Brooks (1997) manipulated depth of processing in their first experiment. They obtained a mirror effect in both recognition and classification tasks. Increasing the depth of processing from shallow to deep resulted in correct endorsements (HRs) increasing and inappropriate new word endorsements decreasing (FAs). Stretch and Wixted (1998) induced a mirror effect by manipulating the number of times words were displayed in the study phase. One group of participants saw each word once (weak condition) and another saw each word three times (strong condition). Both manipulations were designed to increase the strength of evidence for old words. Jacoby (1999) also used repetition at study to increase the strength of familiarity. Hence manipulating study strength in a similar way should produce both a mirror effect and an increase in familiarity. The experiments in this chapter thus used a study-strength manipulation. In order to ensure participants paid attention, the meaning rating task of Higham and Brooks' deep-processing condition was used.

The following experiments used a similar procedure to the second experiment from Chapter 2 (see section 2.3.3). Stretch and Wixted (1998) and Higham and Brooks (1997) used single words at test rather than word pairs. In order to bring Experiment 3 into alignment with these studies, single words were presented at test. A second benefit of this approach concerns the number of data points that contribute to the analysis. Signal-detection analysis on word pairs requires the use of two words to produce one data point (Macmillan & Creelman, 2005). Using single words at test rather than word pairs thus doubles the number of data points obtained from the same number of stimuli. As in Chapter 2, confidence ratings,

phenomenological ratings and questionnaires were used to judge participants' awareness of the knowledge they were using to discriminate the words.

3.1.2 Predictions.

Stretch and Wixted (1998) concluded that a strength-based mirror effect occurred due to a criterion shift. The following predictions will be based on signal-detection model assuming three distributions placed along a strength-of-evidence dimension – one for NI words, one for NC and one for old words. A word is endorsed as old in recognition or rule consistent in classification if the observed strength of evidence is above a criterion. The term "endorsements" will be used to refer to any word endorsed as old in recognition or rule consistent in classification.

First, consider how study strength may affect recognition performance as modelled by SDT. The strength of evidence of old items in the strong-study (each word presented three times) condition would be greater than in the weak-study (each word presented once) condition, shifting the old distribution to the right. The NI item distribution should be unaffected by the study-strength manipulation as they were not studied and share no rule consistent features with the old items. There is some question about what may happen with the NC items. One possibility is that the NC distribution would stay static as NC items are not strengthened directly. Alternatively, because old and NC items share the features of rule consistence, strengthening old items may result in a residual strengthening of NC items.

The study-strength manipulation might also affect the placement of the old/new decision criterion. Because the strength of evidence associated with the old distribution is greater in the strong-study rather than the weak-study condition, participants may adopt a more conservative criterion. This would reduce inappropriate NI endorsements resulting in greater accuracy in strong-study rather than weak-study conditions. Coupled with an increase in appropriate old endorsements, this would result in a strength-based mirror effect. Alternatively, participants might not respond to the study-strength manipulation in which case no criterion shift will occur.

Thus a between-list manipulation of study strength could affect recognition performance by shifting the NC distribution (in addition to the old distribution), by shifting the criterion, or both. The endorsement rate changes for each of the four possible scenarios are outlined in Table 3.1 and Figure 3.1. This assumes that should a criterion shift occur, it would be the same as any shift in the NC distribution but not as great as the shift in the old

distribution. There are two possible mirror effects that could occur (see Table 3.1). A mirror effect could occur with respect to NC and NI items. This is reflected in Case 1. If the NC distribution shifts in lock-step with the criterion shift, then there will be a mirror effect only involving NI items (Case 2). Cases 3 and 4 reflect the lack of a mirror effect. A criterion shift will result in a mirror effect with at least NI items and possibly NC items. Other patterns are possible, but Table 3.1 represents the most likely set of possibilities.

Table 3.1

Predicted Changes in Recognition Endorsement Rates from Weak-study to Strong-study Conditions

	With Criterion Shift		Without Criterion Shift	
Word type	Case 1 Without NC distribution shift	Case 2 With NC distribution shift	Case 3 Without NC distribution shift	Case 4 With NC distribution shift
Old	Increase[1,2]	Increase[3]	Increase	Increase
NC	Decrease[1]	No change	No change	Increase
NI	Decrease[2]	Decrease[3]	No change	No change

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

[1,2,3] = Pairs of changes that interact to produce a strength-based mirror effect.

The effect of study strength in the classification task is more difficult to predict. Distribution shifts, criterion shifts and endorsement rate changes are possible in the same manner as in the recognition task (see Table 3.1). NC endorsements are appropriate in the classification task where they are not in recognition. This may result in participants adopting a more liberal criterion in classification, resulting in higher old, NC and NI endorsements in classification than in recognition. However, Higham and Brooks (1997) found similar old endorsement rates in their classification and recognition tasks. If participants adopted a more liberal criterion in classification than in recognition but old endorsement rates did not change, this would suggest that the old distribution shifts to the left from recognition to classification. One possible explanation for this is that participants rely on recollection less in classification than in recognition, resulting in some of the old words being missed. If this occurs, then in this experiment the same pattern of changes will be observed in classification and recognition, but with classification having higher overall endorsement rates for NC and NI and similar endorsement rates for old items compared to recognition.

Higham, Bruno and Perfect (2010) demonstrated that test words that have a high contextual similarity to words on the study list can induce a reversal of the FA portion of the mirror effect. However, in Experiment 1 of that paper, Higham et al. used the same stimuli as in the following experiments and did not find a reversal of the mirror effect due to the rule set. Consequently, no mirror-effect reversal was expected here.

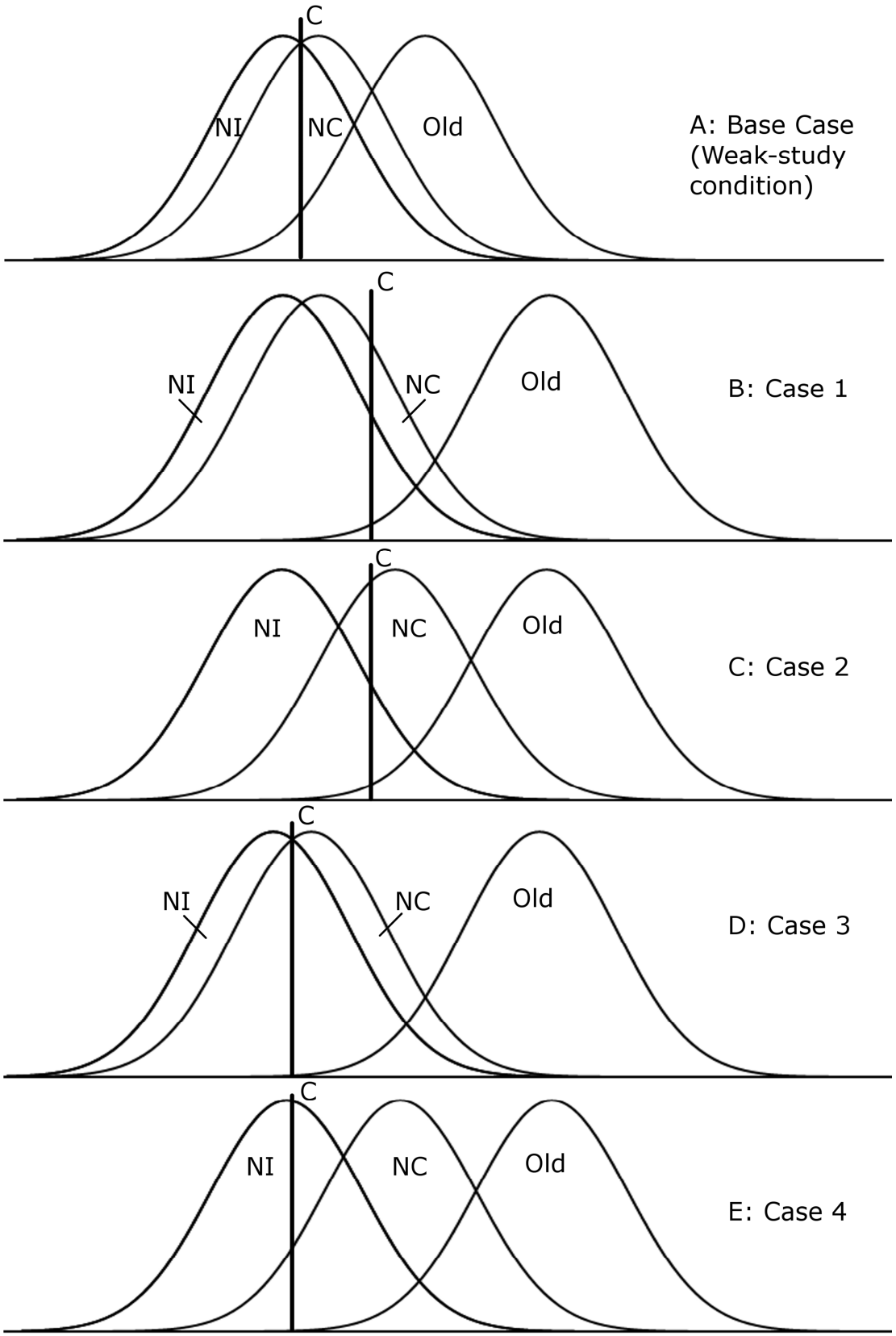

Figure 3.1. Distribution and criterion shifts that could occur as a result of a study-strength manipulation - Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words, C = Criterion.

3.2 Experiment 3

3.2.1 Method.

3.2.1.1 Participants.

Forty-two undergraduates from the UoS participated in the experiment. All participants were given either course credits or £5 payment for their time.

3.2.1.2 Materials and design.

Two study lists of 80 words and four test lists of 120 words each were used. The study lists (Study List A and Study List B) were created in the same way as in Experiment 1. A study phase was administered using a study-strength manipulation. Half of the participants saw each study word once (weak-study condition) whilst the other half of the participants saw each study word three times (strong-study condition). The study list used was counter-balanced across all participants such that each study list was used equally in the weak-study and strong-study conditions. The word presentation order was randomised for each participant, with the restriction that all words on a study list had to be seen once before any word could be repeated.

Two test lists were created to go with each study list – Test List 1 and Test List 2 went with Study List A and Test List 3 and Test List 4 went with Study List B. First 40 new words of each type were randomly chosen from the word pools used in Experiments 1 and 2 – that is 40 CC words, 40 RA words, 40 RC words and 40 CA words. Twenty words of each type were randomly selected to appear on Test Lists 1 and 3 and the other twenty words of each type appeared on Test Lists 2 and 4. Eighty words from Study List A (old words) were then equally split between Test Lists 1 and 2 ensuring that equal numbers of each word type (20 CC and 20 RA) appeared on each test list. Eighty words from Study List B were split between Test Lists 3 and 4 in the same manner. Thus Test Lists 1 and 3 both contained the same set of 80 new words but different sets of 40 old words, and the same was true of Test Lists 2 and 4. The final test lists therefore contained 40 old words, 40 NC words and 40 NI words each, with the NC words from Test Lists 1 and 2 acting as the NI words on Test Lists 3 and 4 respectively (and vice versa). Once created, each list was not changed, although the presentation order of the words was randomised anew for each participant.

At test, participants had to complete two tasks. A recognition task involved participants making old/new judgments to individually presented words and a classification

task involved participants making consistent/inconsistent judgments to individually presented words. Half the participants completed the recognition task first whilst the other half completed the classification task first. Participants that had seen Study List A were given Test Lists 1 and 2 whilst participants that had seen Study List B were given Test Lists 3 and 4. Assignment of test list to task was counterbalanced across participants. For both tasks participants were asked to make their recognition or classification decision, give a confidence rating and then give an attribution just as in Experiment 2. Thus the experiment used a 2 (study strength) x 2 (task type) x 3 (word type) design. Finally, a questionnaire was used to assess verbalisable knowledge of the rule set – this was the same questionnaire used in Experiment 2 except participants were given both Questions 1a and 1b (see Appendix B).

3.2.1.3 Procedure.

After consenting, all participants completed the experiment on an Apple Macintosh computer. The study phase consisted of 80 words from one of the two study lists. Each word was displayed in the centre of the screen for 1.5 seconds. Presentation order of the words was randomised separately for each participant. Attention was ensured by participants rating each word for how well they understood its meaning on a scale of one to four where one indicated that they did not understand the meaning of the word at all whilst four indicated that they fully understood the meaning of the word. In the weak-study condition, each word was displayed once and in the strong-study condition each word was displayed three times to make a total of 80 word presentations in the weak-study condition and 240 presentations in the strong-study condition. Participants were not told about the rule set at this stage.

After the study phase, participants were given two test phases on the computer – one contained a yes/no recognition test and the other contained a yes/no classification test. Assignment of test list to task type was counterbalanced and the presentation order of the words was random. In the recognition task, participants judged each word as either old (having been seen before) or new (not seen before) by clicking a radio button under the word. They also typed in a confidence rating on a scale of 50-100 and selected the basis of their decision from a choice of random chance, intuition, memory and rules using a radio button. The classification task was the same as the recognition task except that participants judged words for their rule consistence instead of the old/new decision. All participants took both recognition and classification tests, half took the recognition task first and half took the classification task first. The test was self-paced and all decisions were made on the same screen.

Once both test phases had been completed a computer-based questionnaire was administered in order to probe for understanding of the rule set (see Appendix B).

3.2.2 Results.

For both the recognition and classification tasks there were endorsement rates associated with old words, NC words and NI words. Two types of *d'* were calculated[4]. Following the language of Higham and Brooks (1997), the episodic *d'* represents the ability of participants to discriminate between old and NC items and reflects the contribution of remembering the stimuli from test over and above that of the rule consistence status of the item. The structural *d'* represented the ability of participants to discriminate between NC and NI words and reflects the contribution of the rule consistence without the influence of veridical episodic memory for items. Endorsement rates are presented as proportions, although proportions of zero or one were corrected as recommended by Macmillan and Creelman (2005) by replacing rates of 1 with 1-(1/2n) and rates of 0 with 0 + (1/2n) where n was the total number of possible endorsements (n = 40 for most of the analysis). This correction was applied twenty two times, most frequently on old endorsement rates that were at ceiling.

Please also note that, as in Chapter 2, where *t* tests are calculated the effect size *d* is also reported – this is not to be confused with the measure of discrimination *d'*. There were no interactions involving test list order or the exact study list used so these variables are not addressed further. All interactions were investigated using Bonforonni corrected pairwise comparisons. Some tables include "overall" columns – these are presented for clarity where an interaction resulted in the data being collapsed across the relevant variable.

3.2.2.1 Analysis of episodic and structural d'.

Initially the episodic and structural *d'* were analysed in order to investigate whether participants used the episodic and structural status of stimuli to make their decisions. The presence of an effect is indicated by *d'* being higher than zero. This was tested by seeing if the lower bound of 1.96 standard errors (i.e. the 95% confidence interval) for each *d'* intersected zero. No intersection with zero meant that the relevant *d'* represented above-

[4] Episodic *d'* was calculated by subtracting the z transformed NC endorsement rate from the z old endorsement rate. Structural *d'* was calculated by subtracting the z NI rate from the z NC rate. Traditional SDT calculates *d'* using hit and false-alarm rates. However, since the NC rate is a false-alarm rate in recognition and a hit rate in classification this terminology has been avoided.

chance discrimination (see Table 3.2*)*. Participants used both the structural and episodic status of the stimuli to make decisions in both tasks.

Table 3.2

d' by Study Strength, Effect Type and Task Type in Experiment 3 (SE in brackets)

Task and *d'* type	Weak study	Strong study	Total
Recognition			
Episodic *d'*	2.36 (.14)*	2.80 (.18)*	2.58 (.11)
Structural *d'*	.30 (.10)*	.24 (.06)*	.27 (.06)
Classification			
Episodic *d'*	1.88 (.18)*	2.29 (.11)*	2.09 (.11)
Structural *d'*	.21 (.09)*	.30 (.09)*	.26 (.06)
Both tasks			
Episodic *d'*	2.12 (.13)	2.55 (.13)	2.33 (.09)
Structural *d'*	.25 (.07)	.27 (.07)	.26 (.05)

* = Lower bound of 95% confidence interval above chance level of 0.

The sensitivity of the *d'* measures to study strength was investigated with a 2 x 2 x 2 mixed ANOVA with task type (recognition or classification) and effect type (episodic or structural) as within-subject variables and study strength (weak study or strong study) as a between-subject variable. See Table 3.3 for the results of the ANOVA as well as related pairwise comparisons.

Table 3.3

Results of 2 (Study Strength) x 2 (Task Type) x 2 (Effect Type) ANOVA on d' in Experiment 3

Effect	*F* value (df)	*p* value	η^2
Task type	$F(1, 41) = 13.28$	.001*	.25
Effect type	$F(1, 41) = 413.16$	< .001*	.91
Study strength	$F(1, 41) = 4.99$	.03*	.11
Study strength by effect type	$F(1, 41) = 4.09$	.05*	.09
Study strength for episodic d'	$F(1, 41) = 5.80$	.02*	.13
Study strength for structural d'	$F < 1$	-	-
Task type by effect type	$F(1, 41) = 8.53$	.006*	.18
Task type for episodic d'	$F(1, 41) = 13.67$	.001*	.25
Task type for structural d'	$F < 1$	-	-
There were no other significant effects	$F < 1$	-	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

The main effect of task type indicated that there was better discrimination in recognition ($M = 1.43$, $SE = .06$) than classification ($M = 1.17$, $SE = .06$). The main effect of effect type indicated a higher episodic *d'* ($M = 2.35$, $SE = .09$) than structural *d'* ($M = .26$, $SE = .05$). The main effect of study strength indicated better discrimination in strong-study ($M = 1.41$, $SE = .07$) than weak-study ($M = 1.19$, $SE = .07$) conditions.

The interaction between effect type and task type was due to the episodic *d'* being lower in classification than in recognition, whilst the structural *d'* was unchanged between classification and recognition (see Table 3.2 for means and standard errors). The interaction between effect type and study strength reflected the fact that the episodic *d'* was greater in strong-study rather than weak-study conditions, whilst the structural *d'* was unchanged between weak-study and strong-study conditions (see Figure 3.2 and Table 3.2).

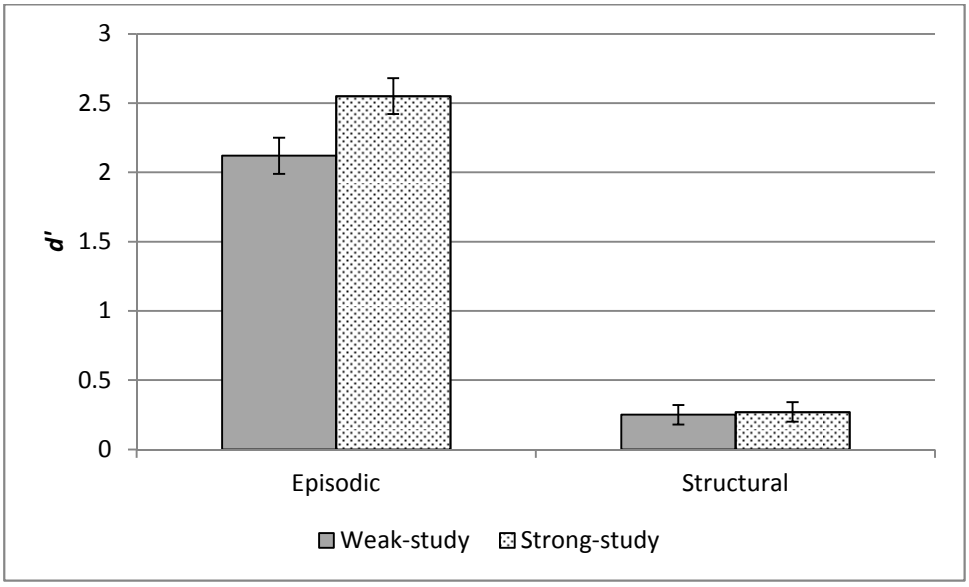

Figure 3.2. Episodic and structural *d'* by study strength in Experiment 3.

3.2.2.2 Analysis of endorsement rates.

Endorsement rates (see Table 3.4) were entered into a 3 x 2 x 2 ANOVA with word type (old, NC and NI), study strength (weak study or strong study) and task type (recognition or classification). See Table 3.5 for the results of the ANOVA and related pairwise comparisons.

Table 3.4

Endorsement Rates by Study Strength, Task Type and Word Type in Experiment 3 (SE in brackets)

Task and word type	Weak study	Strong study	Total
Recognition			
Old	.88 (.02)	.92 (.02)	.90 (.02)
NC	.18 (.03)	.15 (.02)	.16 (.02)
NI	.13 (.03)	.11 (.02)	.12 (.02)
Classification			
Old	.87 (.02)	.93 (.02)	.90 (.02)
NC	.32 (.04)	.31 (.04)	.32 (.03)
NI	.26 (.04)	.22 (.03)	.23 (.03)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 3.5

Results of 2 (Study Strength) x 2 (Task Type) x 2 (Word Type) ANOVA on Endorsement Rates in Experiment 3

Effect	*F* value (df)	*p* value	η^2
Task type	$F(1, 40) = 20.55$	< .001*	.34
Word type	$F(2, 80) = 1041.05$	< .001*	.96
Old versus NC	$F(1, 41) = 1055.84$	< .001*	.96
NC versus NI	$F(1, 41) = 26.70$	< .001*	.39
Task type by word type	$F(2, 80) = 15.82$	< .001*	.28
Task-type for old words	$F < 1$		
Task-type for NC words	$F(1, 40) = 22.43$	< .001*	.36
Task-type for NI words	$F(1, 40) = 17.95$	< .001*	.31
Word type by study strength – marginal interaction (no pairwise conducted)	$F(2, 80) = 3.25$	< .06	.07
There were no other significant effects	$F < 1$		

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction. * = *p* value denotes significant difference.

The main effect of task type reflected more endorsements in classification (M = .49, SE = .02) than in recognition (*M* = .39, *SE* = .01). The main effect of word type reflected more old endorsements (*M* = .90, *SE* = .01) than NC endorsements (*M* = .24, *SE* = .01), and more NC endorsements than NI endorsements (*M* = .18, *SE* = .02).

The interaction reflected the fact that old endorsements did not differ from classification to recognition, whilst there were more NC and NI endorsements in classification than in recognition (see rightmost column in Table 3.4).

Overall, the failure to observe any effect of study strength means that none of the predicted patterns were observed, as they all rest on such an effect being obtained. The observed pattern was closest to the case where there is no criterion shift and no NC distribution shift. Inspection of the endorsements rates reveals a possible reason for the lack of study-strength effect – old endorsements are near ceiling in the weak-study condition and so an increase in old endorsement rates by study strength may not have been possible. The marginal interaction of word type and study condition hints at a mirror effect, and explains why the episodic *d'* increased from weak-study to strong-study conditions. Old endorsements descriptively increased and both NC and NI endorsement descriptively

decreased enough to produce the change in the episodic *d'*, but not enough to produce an interaction in the ANOVA.

3.2.2.3 Performance by attribution.

It is not clear what meaning a *d'* measure would have if calculated for each attribution type, as attribution decisions could take the form of either criteria above and below which different attributions are made, or the contributions of different underlying processes. Thus endorsement rates were broken down into attribution types – for instance endorsements for old stimuli in recognition were split down according to which attribution each endorsement was given. In Chapter 2 guess responses were mostly associated with chance performance, and guess responses are often excluded from such analyses (Higham et al., 2010). Consequently, guess responses were excluded from the following analysis. Rules responses were infrequently used and were excluded. Intuition and memory responses were each entered into 3 x 2 x 2 repeated measures ANOVA with word type (old, NC and NI), study strength (weak study versus strong study) and task type (recognition versus classification) – see Table 3.6 for means and standard errors and Table 3.7 for the ANOVA results.

Table 3.6

Mean Endorsement Rates by Task Type, Word Type, Study Strength and Attribution in Experiment 3 (SE in brackets)

Task and word type	Weak study		Strong study	
	Intuition	Memory	Intuition	Memory
Recognition				
Old	.12 (.02)	.64 (.05)	.09 (.02)	.79 (.03)
NC	.08 (.01)	.06 (.02)	.07 (.01)	.04 (.01)
NI	.07 (.02)	.02 (.01)	.04 (.01)	.03 (.01)
Classification				
Old	.10 (.02)	.58 (.06)	.10 (.02)	.75 (.05)
NC	.13 (.02)	.05 (.01)	.15 (.03)	.06 (.01)
NI	.09 (.02)	.03 (.01)	.13 (.03)	.03 (.01)
Both Tasks				
Old	.12 (.02)	.64 (.05)	.09 (.02)	.79 (.03)
NC	.08 (.01)	.06 (.02)	.07 (.01)	.04 (.01)
NI	.07 (.02)	.02 (.01)	.04 (.01)	.03 (.01)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 3.7

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) ANOVAs on Intuition and Memory Attributions in Experiment 3

Effect	*F* value (df)	*p* value	η^2
Intuition ANOVA			
Task type	$F(1, 40) = 6.66$	.01*	.14
Task type by word type	$F(2, 80) = 7.78$	.001*	.16
Task type for old words	$F < 1$	-	-
Task type for NC words	$F(1, 40) = 8.27$	.01*	.17
Task type for NI words	$F(1, 40) = 8.06$	.01*	.17
There were no other significant effects	$F(2, 80) = 2.28$	.11	-
Memory ANOVA			
Word type	$F(2, 80) = 417.93$	< .001*	.89
Old versus NC	$F(1, 41) = 362.09$	< .001*	.90
NC versus NI	$F(1, 41) = 17.76$	< .001*	.30
Study strength	$F(1, 40) = 4.61$	.04*	.08
Word type by study strength	$F(2, 80) = 6.19$	.003*	.13
Study strength for old words	$F(1, 40) = 6.02$	.02*	.13
Study strength for NC words	$F < 1$	-	-
Study strength for NI words	$F < 1$	-	-
There were no other significant effects	$F(2, 80) = 3.13$	.08	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

For intuition, the task-type main effect reflected more intuition attributions in classification ($M = .13$, $SE = .01$) than in recognition ($M = .08$, $SE = .01$). The interaction was further investigated with Bonferroni corrected pairwise comparisons. It reflected the fact that there were the same intuition attributions for old words in both recognition and classification whilst there were more intuition attributions for NC and NI words in classification than in recognition.

The memory main-effect of word type reflected more memory attributions for old words ($M = .69$, $SE = .03$) than NC ($M = .05$, $SE = .01$) and more NC attributions than NI ($M = .03$, $SE = .01$). The study-strength main effect reflected more memory attributions in strong-study ($M = .28$, $SE = .02$) than in weak-study ($M = .23$, $SE = .02$) conditions. The

word-type and study-strength interaction reflected the fact that memory attributions only increased by study strength for old words whilst NC and NI memory attributions stayed the same (see Table 3.6 for means).

3.2.2.4 Awareness measures.

3.2.2.4.1 Awareness – questionnaire.

The questionnaire indicated that no participant developed knowledge of the rule set. Even when directly told that the rule was a conjunctive rule and given a list of possible rules, no participant described the actual rule set.

3.2.2.4.2 Awareness - guessing criterion.

As a *d'* measure would be hard to interpret for the guessing criterion (see section 3.2.2.3) no guessing criterion tests were used here or for the rest of the experiments in this thesis. It is worth noting that an ANOVA on guess responses revealed no differences by any variable, highest $F(1, 40) = 3.34$, $p = .07$. This suggests that people were using this category appropriately for chance-level responses only.

3.2.2.4.3 Awareness – zero-correlation criterion.

Awareness was tested with the zero-correlation criterion. If confidence for correct answers was different from the confidence for incorrect answers then explicit awareness is implicated. In this case responses associated with the episodic *d'* and the structural *d'* were tested against the zero-correlation criterion. For the episodic *d'* confidence for correct answers to old and NC stimuli were compared to confidence for incorrect answers to old and NC stimuli. For the structural *d'* confidence for correct and incorrect answers for NC and NI stimuli were used. The only issue of interest is whether the correct confidence is higher than the incorrect confidence in each condition. Thus for simplicity confidences were compared with *t* tests – see Table 3.8 for means and standard errors.

Table 3.8
Mean Confidence for the Episodic and Structural d' Measures by Study Strength and Task Type (SE in brackets)

Task and effect type	Weak study		Strong study	
	Correct	Incorrect	Correct	Incorrect
Recognition				
Episodic	86.07 (1.60)	70.08 (2.37)	88.41 (1.82)	69.90 (3.01)
Structural	81.01 (2.06)	70.47 (2.43)	84.32 (2.19)	67.81 (2.95)
Classification				
Episodic	78.47 (2.05)	66.71 (2.83)	81.38 (2.16)	70.00 (3.08)
Structural	69.04 (2.36)	68.90 (2.47)	71.40 (1.90)	69.26 (2.24)

In recognition, mean correct confidence was higher than mean incorrect confidence for the episodic *d'* in both weak-study, $t(16) = 8.03$, $p < .001$, $d = 1.58$, and strong-study conditions, $t(15) = 7.90$, $p < .001$, $d = 1.36$. For the structural *d'*, mean correct confidence was higher than incorrect confidence in weak-study, $t(16) = 4.97$, $p < .001$, $d = 1.07$ and in strong-study conditions, $t(15) = 5.72$, $p < .001$, $d = 1.12$.

In classification, correct confidence was higher than incorrect confidence for the episodic *d'* in both weak-study, $t(17) = 6.44$, $p < .001$, $d = 1.11$, and strong-study conditions, $t(13) = 5.88$, $p < .001$, $d = 0.97$. On the other hand correct confidence was not higher than incorrect confidence for the structural *d'* in either weak-study or strong-study conditions, highest $t(19) = 2.08$, $p = .05$. See Table 3.9 for a summary of the results.

Table 3.9
Zero-correlation Awareness Summary for Experiment 3

Task and *d'* type	Weak study	Strong study
Recognition		
Episodic	Aware	Aware
Structural	Aware	Aware
Classification		
Episodic	Aware	Aware
Structural	Unaware	Unaware

3.2.3 Discussion.

Consistent with Higham and Brooks (1997), there was evidence of both a structural and an episodic effect in both recognition and classification tasks. Additionally, the episodic effect was larger than the structural effect in all conditions. The presence of an episodic effect in both recognition and classification tasks suggests that participants used the episodic status of the word to help identify rule-consistent words. The presence of a structural effect in the recognition task on the other hand suggests that participants inappropriately endorsed NC words more often than NI words, indicating that the rule-consistence status of the word resulted in a feeling, such as familiarity, which participants misinterpreted as diagnostic of the episodic status of a word. The study-strength manipulation increased the magnitude of the episodic effect whilst leaving the structural effect untouched. This could simply be because of floor effects on NC and NI endorsement rates and a ceiling effect on the old endorsement rate. Further discussion of this result is reserved for Experiments 4 and 5, in which the floor and ceiling effects were addressed.

The patterns of endorsement rates involving the study-strength manipulation were not consistent with any of the predicted scenarios. Within recognition, endorsement rates were similar in weak-study and strong-study conditions. A ceiling effect may have prevented a significant increase in old endorsement rates. If the expected increase in the old endorsement rates from weak-study to strong-study conditions had occurred, then the conclusion would have been that neither the criterion nor the NC distribution shifted with study strength (as in Case 3 in the predictions). Old word endorsements did increase by study strength in the memory attribution category. If the memory category at least partly represents recollection, then this suggests that the study-strength manipulation increased recollection to some extent. This seems to have resulted in only a small increase in the overall old endorsement rate, partly because it was accompanied by a small descriptive decrease in the old endorsement rate for intuition attributions. Increases in study strength usually result in an increase in both recollection and familiarity (Jacoby, 1999). It may be that because old endorsements were already close to ceiling the study-strength manipulation had the effect of converting some familiarity based responses to recollection-based responses, and thus converting some intuition old endorsements to memory old endorsements. At the least, the presence of the old-endorsement portion of a mirror effect in the memory attributions suggests that part of the mirror effect may be present but masked by the high overall old endorsements.

Participants did not shift their criterion. A lack of a criterion shift has been demonstrated for within-list manipulations under certain conditions (Bruno et al., 2009; Stretch & Wixted, 1998), but between-list manipulations usually find a criterion shift (Hockley & Niewiadomski, 2007; Stretch & Wixted, 1998). Although some doubts have been expressed over whether participants use strength to set their criterion, Verde and Rotello (2007) demonstrated that participants set their criterion according to strength in the first block of an experiment and then did not change it. The results obtained here would suggest that participants do not seem to take account of strength when setting their criterion. However, if participants set their criterion in response to the perceived strength of the old words, then the fact that old endorsements are at ceiling may result in a similar criterion being set by weak-study and strong-study participants. In order to investigate this possibility, Experiment 4 attempted to remedy the old endorsement ceiling effect.

The pattern of endorsement changes from recognition to classification are more consistent with predictions - both NC and NI endorsements rates increased whilst old endorsements stayed static. There are two possible factors that would produce this pattern of changes (see Figure 3.3 below). On the one hand (Panel B - Case 1), the NC and NI distributions could be higher on the strength-of-evidence scale in classification than in recognition whilst the criterion and old distributions are static. On the other hand (Panel C - Case 2), the old distribution could be lower on the strength-of-evidence scale in classification than in recognition accompanied by a downwards criterion shift, whilst NC and NI stay static. Changes in the distribution positions between recognition and classification could indicate either different processes contributing to strength of evidence in each task, or else some other factor resulting in different strengths of evidence across the tasks. Although the data does not favour either of these explanations, it is logical to assume that in the classification task participants are not attending as closely to old words at test, and thus may not be experiencing the maximum strength of evidence that old words could produce. In other words, because they are not specifically looking for old words, some old words are missed. Such an explanation is consistent with the idea that recollection is a deliberate process that can be applied or not applied depending on perceived task demands. An alternative theory is that the conscious effort to look for rules in the classification condition brings in evidence that would otherwise be ignored in the recognition condition, increasing the strength of evidence of the NC distribution (as in Case 1). However, it is difficult to see why this additional evidence would also increase the strength of the NI distribution along with the NC distribution. The additional evidence would either be appropriate, in which case the NC

distribution would shift to the right more than the NI distribution, or else it would be inappropriate, in which case the gap between the NC and NI distributions would decrease. As not one participant evidenced verbalisable knowledge of the rules, it is also questionable what additional evidence could be brought to bear by conscious effort. Given these factors, a criterion shift along with an old distribution shift is the more parsimonious explanation for changes in endorsement rates between recognition and classification. However, this remains as an avenue of future investigation.

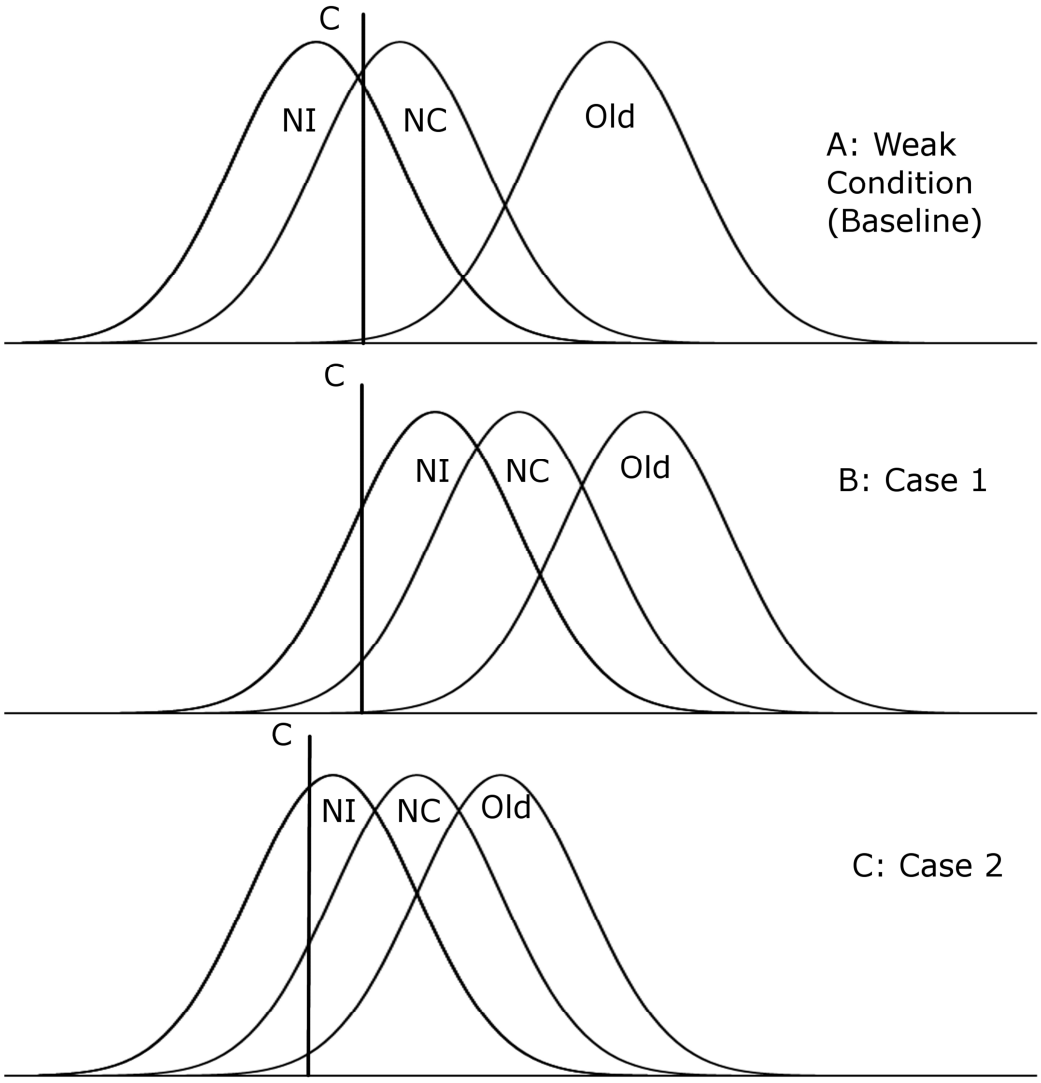

Figure 3.3. Possible reasons for differences in endorsements by task.

The evidence from the various measures of awareness suggested that no participant gained explicit knowledge of the rule set. Guess attributions were insensitive to changes in all variables, suggesting that the guess category really was being used for guesses. The zero-

correlation criterion indicated that participants could tell the difference between their episodic correct and incorrect answers in both classification and recognition suggesting that participants were aware of the episodic status of the items. For the structural effect, participants assigned higher confidence to correct answers in recognition but there was no difference in classification. So in recognition, when participants inappropriately endorse a word they do so with low confidence, whilst when they appropriately reject a rule-consistent word they do so with high confidence. This is consistent with Scott & Dienes (2008) "calibrated familiarity model" (CFM) in which participants expect a mean level of familiarity and then use deviations from this mean level to set confidence and make decisions. This way, a rule-consistent word that evinces a similar level of familiarity to an old word with low familiarity could be endorsed with low confidence, whilst rule-consistent words that have even lower levels of familiarity might be rejected with high confidence. Why confidence for correct answers would be higher than confidence for incorrect answers in recognition but not classification is not clear. Participants may consciously experience differences in familiarity in both recognition and classification but have different attitudes to the different tasks. In recognition, they may not be aware of a rule set existing and so be willing to express confidence in low levels of familiarity that they used to endorse a word. In classification, they experienced the same feeling of familiarity but may not have realised it could be used to tell rule-consistent from rule-inconsistent words, and thus did not take it into account when making their metacognitive confidence judgements. A second possibility is that participants made confidence decisions in the way that Scott and Dienes suggested and are not as well calibrated to familiarity in the classification condition. This would result in smaller amounts of familiarity being given higher confidence in the better calibrated recognition task than in the classification task. This could happen if participants have greater amounts of pre-experimental practice with recognition-type decisions and in expressing confidence in those decisions. Either way, no participant described the rules in the questionnaire. Taken together with the presence of a structural effect, this is strong evidence that participants were sensitive to rule consistence but not able to explicitly explain the rule set. This result is fairly typical of many experiments involving implicit learning (Higham & Brooks, 1997; A. S. Reber, 1989).

Overall, Experiment 3 was unsuccessful in obtaining a mirror effect because of the ceiling effect for old endorsements. Experiment 4 impaired learning conditions relative to Experiment 3 in order to reduce performance based on memory for studied words, and therefore old endorsements.

3.3 Experiment 4

3.3.1 Method.

3.3.1.1 Participants.

Sixty-four undergraduates from the UoS participated in the experiment. All participants were either given course credits or £5 payment.

3.3.1.2 Materials and design.

The materials and general design were the same as in Experiment 3.

3.3.1.3 Procedure.

The procedure was identical to Experiment 3, except for the following changes designed to address the ceiling effect for old endorsements. The words in the study phase were displayed for 1 second each rather than 1.5 seconds. Participants did not have to rate the words for understanding. Thus, where Experiment 3's study phase was self-paced and allowed some time for rehearsal and reflection on each word, Experiment 4 was fixed pace with little time for rehearsal.

3.3.2 Results.

Endorsement rates were calculated and corrected as in Experiment 3. Eight participants were excluded from the analysis – seven did not follow the instructions or guessed on every trial and one displayed low recognition performance (more than 3 standard deviations below the mean).

3.3.2.1 Task order.

Although there were no effects involving the order in which the tasks were given, a visual inspection of the old word endorsement rates indicated that for the recognition task only the weak-study endorsement rates seemed to differ by task order. Subjecting just these rates to a *t* test revealed that there were more endorsements for old words ($M = .71$, $SE = .04$) when the recognition test was given first than when the recognition task was given second ($M = .53$ $SE = .06$), $t(20) = 2.63$, $p = .02$, $d = 1.00$. As it is possible that this could have a large effect on the data, this interference effect was avoided by only analysing the first phase of each participant. Thus task type becomes a between-subjects variable rather than a within-subjects variable.

3.3.2.2 Analysis of episodic and structural d'.

The episodic and structural *d'* measures were again compared to chance using confidence intervals and the changes in magnitude of the *d'* measures were investigated with a 2 x 2 x 2 ANOVA with task type, effect type and study strength - see Table 3.10 for means and standard errors and Table 3.11 for the ANOVA results.

Table 3.10

d' by Study Strength, Effect Type and Task Type from Experiment 4 (SE in brackets)

Task and *d'* type	Weak study	Strong study
Recognition		
Episodic *d'*	1.04 (.10)*	1.54 (.11)*
Structural *d'*	.29 (.07)*	.35 (.09)*
Classification		
Episodic *d'*	.76 (.10)*	1.20 (.15)*
Structural *d'*	.17 (.11)	.14 (.08)
Both Tasks		
Episodic *d'*	.91 (.07)	1.38 (.09)
Structural *d'*	.23 (.06)	.25 (.06)

* = Lower bound of 95% confidence interval above chance level of 0.

Table 3.11

Results of 2 (Study Strength) x 2 (Task Type) x 2 (Effect Type) ANOVA on d' in Experiment 4

Effect	*F* value (df)	*p* value	η^2
Task type	$F(1, 52) = 13.16$	< .001*	.20
Effect type	$F(1, 52) = 131.86$	< .001*	.72
Study strength	$F(1, 52) = 14.04$	< .001*	.90
Effect type by study strength	$F(1, 52) = 8.51$	.005*	.14
Study strength for episodic *d'*	$F(1, 52) = 17.07$	< .001*	.25
Study strength for structural *d'*	$F < 1$	-	-
There were no other significant effects	$F < 1$	-	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

The effect-type main effect reflected a higher episodic *d'* (*M* = 1.14, *SE* = .06) than structural *d'* (*M* = .24, *SE* = .04). The study-strength main effect reflected better discrimination in the strong-study condition (*M* = .81, *SE* = .05) than the weak-study condition (*M* = .57, *SE* = .05). The task-type main effect represented better discrimination in the recognition task (*M* = .81, *SE* = .04) than the classification task (*M* = .57, *SE* = .05).

The effect-type by study-strength interaction was further investigated with pairwise comparisons (see Table 3.10 for means and standard errors). As in Experiment 3, the episodic *d'* was greater in strong-study rather than weak-study conditions, whilst the structural *d'* did not change from weak-study to strong-study conditions (See Figure 3.4).

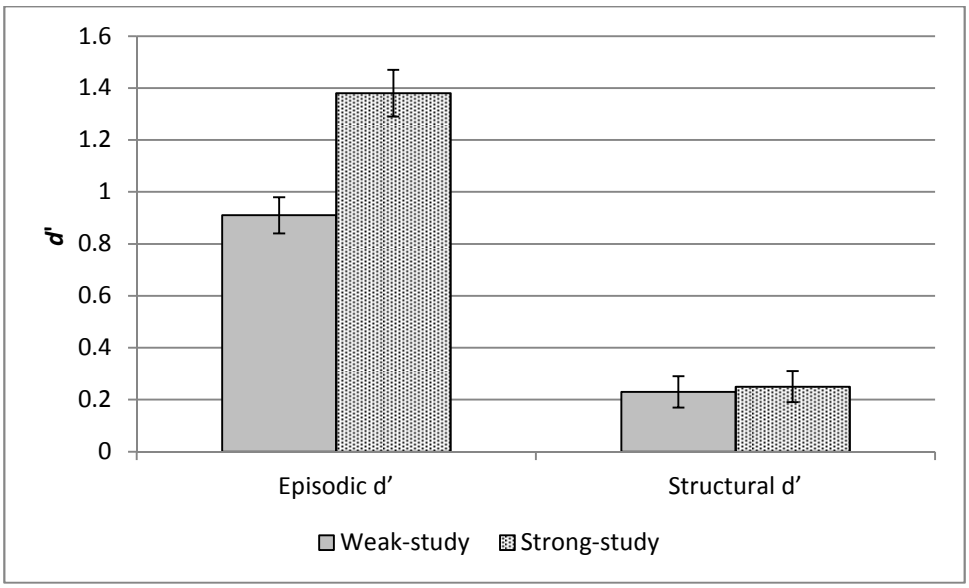

Figure 3.4. Episodic and structural effects by study strength in Experiment 4

3.3.2.3 Analysis of endorsement rates.

Endorsement rates were again compared across word type, study strength and task types, see Table 3.12 for means and standard errors and Table 3.13 for the results of the ANOVA.

Table 3.12

Endorsement Rates by Study Strength, Word Type and Task Type from Experiment 4 (SE in brackets)

Task and word type	Weak study	Strong study	Total
Recognition			
Old	.71 (.04)	.80 (.02)	.75 (.02)
NC	.34 (.04)	.26 (.02)	.30 (.03)
NI	.26 (.04)	.18 (.03)	.22 (.03)
Classification			
Old	.67 (.04)	.80 (.04)	.74 (.02)
NC	.40 (.05)	.41 (.04)	.41 (.03)
NI	.33 (.05)	.36 (.04)	.35 (.03)
Both Tasks			
Old	.69 (.02)	.80 (.02)	.75 (.02)
NC	.37 (.03)	.34 (.03)	.35 (.02)
NI	.30 (.03)	.27 (.03)	.28 (.02)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 3.13

Results of 2 (Study Strength) x 2 (Task Type) x 3 (Word Type) ANOVA on Endorsement Rates in Experiment 4

Effect	F value (df)	p value	η^2
Task type	$F(1, 52) = 5.40$	.02 *	.09
Word type	$F(2, 104) = 436.26$	< .001*	.89
Old versus NC	$F(1, 55) = 341.35$	< .001*	.86
NC versus NI	$F(1, 55) = 30.33$	< .001*	.35
Word type by study strength	$F(2,104) = 11.29$	< .001*	.18
Study strength for old words	$F(1, 52) = 11.34$	.001*	.18
Study strength for NC words	$F < 1$	-	-
Study strength for NI words	$F < 1$	-	-
Word type and task type	$F(2, 104) = 11.50$	< .001*	.18
Task type for old words	$F < 1$	-	-
Task type for NC words	$F(1, 52) = 7.52$	.01*	.13
Task type for NI words	$F(1, 52) = 11.48$	.001*	.18
There were no other significant effects	$F(1, 52) = 1.60$	.21	-

Note - Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = p value denotes significant difference.

The main effect of word type reflected more old word endorsements (M = .75, SE = .20) than NC endorsements (M = .35, SE = .02) and more NC endorsements than NI endorsements (M = .28, SE = .02). The main effect of task type reflected more endorsements in classification (M = .50, SE = .02) than in recognition (M = .42, SE = .02).

The interactions were further investigated with pairwise comparisons. The word-type by study-strength interaction (both tasks in Table 3.12) reflected the fact that old endorsements increased from weak-study to strong-study conditions whilst NC and NI endorsements did not change. The word-type by task-type interaction (rightmost column of Table 3.12) reflected the fact that NC and NI endorsements increased from recognition to classification whilst old endorsements remained the same. This was consistent with the pattern seen in Experiment 3 with the addition of an increase in old endorsements, supporting the interpretation that the NC distribution does not shift and there is no criterion shift.

3.3.2.4 Performance by attribution.

Endorsement rates were again broken down into attribution types and were entered into a 3 x 2 x 2 repeated measures ANOVA with word type (old, NC and NI), study strength (weak study vs. strong study) and task type (recognition vs. classification). As in the previous experiment, guess and rules attributions were excluded. See Table 3.14 for means and standard errors and Table 3.15 for the results of the ANOVAs.

Table 3.14

Mean Endorsement Rates by Attribution, Task Type, Word Type and Study Strength (SE in brackets)

Task and word type	Weak study		Strong study	
	Intuition	Memory	Intuition	Memory
Recognition				
Old	.21 (.03)	.40 (.05)	.16 (.03)	.52 (.04)
NC	.17 (.02)	.08 (.02)	.11 (.02)	.07 (.02)
NI	.13 (.02)	.07 (.02)	.08 (.02)	.04 (.02)
Classification				
Old	.17 (.03)	.36 (.05)	.17 (.04)	.48 (.06)
NC	.15 (.02)	.08 (.02)	.15 (.02)	.11 (.02)
NI	.15 (.02)	.09 (.03)	.15 (.02)	.11 (.03)
Both tasks				
Old	.19 (.02)	.38 (.04)	.17 (.02)	.50 (.04)
NC	.16 (.01)	.08 (.01)	.13 (.01)	.09 (.01)
NI	.14 (.01)	.08 (.02)	.12 (.01)	.08 (.02)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 3.15

Results of 2 (Study Strength) x 2 (Task Type) x 3 (Word Type) ANOVA on Intuition and Memory Endorsements in Experiment 4

Effect	*F* value (df)	*p* value	η^2
Intuition ANOVA			
Word type	$F(2, 104) = 7.92$	< .001*	.13
Old versus NI	$F(1, 55) = 12.64$	.001*	.19
Old versus NC/NC versus NI	Highest $F(1, 55) = 5.49$	.02[1]	-
There were no other significant effects	$F(2, 104) = 2.46$	.09	-
Memory ANOVA			
Word-type	$F(2, 104) = 209.46$	< .001*	.80
Old versus NC	$F(1, 55) = 217.50$	< .001*	.80
NC versus NI	$F < 1$	-	-
Word type by study strength	$F(2, 104) = 6.55$	.002*	.11
Study strength for old words	$F(1, 52) = 6.25$	< .001*	.11
Study strength for NC words	*F < 1*	-	-
Study strength for NI words	*F < 1*	-	-
There were no other significant effects	$F(1, 52) = 2.78$	.10	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

1 = Non-significant due to Bonferroni correction.

* = *p* value denotes significant difference.

The intuition word-type main effect reflected more intuition attributions for old words (*M* = .18, *SE* = .02) than for NI words overall (*M* = .13, *SE* = .01). There was no difference between either old or NI intuition attributions and NC intuition attributions (*M* = .15, *SE*=.01)

The memory word-type main effect indicated greater memory attributions for old words (*M* = .44, *SE* = .02) than for NC (*M* = .09, *SE* = .01), and NI words (*M* = .08, *SE* = .01). There was no difference between NC and NI endorsements. Pairwise comparisons for the interaction indicated that memory attributions to old words increased from weak-study to strong-study whilst memory attributions to NC and NI words did not (see Table 3.14).

3.3.2.5 Awareness measures.

3.3.2.5.1 Awareness – questionnaire.

Once again the questionnaire indicated that no participants developed explicit knowledge of the rule set.

3.3.2.5.2 Awareness – zero-correlation criterion.

In order to focus the analysis on the episodic and structural effects the zero-correlation criterion will no longer be calculated or discussed.

3.3.2.6 Combined Analysis of Experiments 3 and 4.

The lack of a mirror effect with a between-list study-strength manipulation was unusual. It would indicate that participants were not shifting their decision criterion in response to the study-strength manipulation. Hockley and Niewiadomski (2007) demonstrated that task complexity can prevent criterion movements. In order to investigate this possibility, a combined analysis of Experiments 3 and 4 was employed. Experiment 3 was defined as 'good' learning conditions (1.5 second exposure, rating task increased depth of processing and rehearsal time) and Experiment 4 was defined as 'poor' learning conditions (1 second exposure, no rating task). Thus a variable of learning conditions was created. Although study time and study task are confounded in this variable, both factors could be said to manipulate depth of processing. If participants did not shift their criterion by depth of processing, then there is corroborating evidence that the task was sufficiently complex to prevent criterion movement. Since Experiment 4 had problems with task order, only the first test phase of each experiment was compared. The analysis focused purely on the mirror effect and therefore only addressed the episodic effect and the endorsement rates that contribute to the effect.

A 2 x 2 x 2 ANOVA on episodic *d'* with between-subjects factors of learning conditions (good versus poor), task type (recognition versus classification) and study strength (weak-study versus strong-study) was conducted in both recognition and classification tasks. Only effects involving learning conditions are reported as other effects repeated the general trends from the two experiments. There was a main effect of learning conditions, $F(1, 79) = 87.62$, $p < .001$, $\eta^2 = 0.53$ and no interactions. The episodic *d'* was higher under good learning conditions ($M = 2.42$, $SE = .10$) than under poor learning conditions ($M = 1.06$, $SE = .10$).

The changes in the episodic *d'* were further investigated with 2 x 2 x 2 x 2 mixed ANOVAs on old and NC endorsement rates. Word type (old vs. NC) was a within-subject factor and learning conditions (good vs. poor), task type (recognition vs. classification) and study strength (weak study vs. strong study) were between-subject factors. Only effects involving learning conditions are reported. There was an interaction between word type and

learning conditions, $F(1, 79) = 62.97$, $p < .001$, $\eta^2 = .44$. Pairwise comparisons indicated that the interaction was due to the old endorsement rate increasing between poor learning conditions ($M = .70$, $SE = .02$) and good learning conditions ($M = .91$, $SE = .02$), $F(1, 79) = 46.80$, $p < .001$, $\eta^2 = .37$, whilst the NC endorsement rate decreased from poor learning conditions ($M = .34$, $SE = .02$) to good learning conditions ($M = .23$, $SE = .02$), $F(1, 79) = 10.20$, $p = .002$, $\eta^2 = .11$.

When looking at the data across learning conditions, the classic mirror effect can be found. The task complexity is not likely to be the reason why the criterion did not shift in response to study strength.

3.3.3 Discussion.

The findings from Experiment 4 should be treated with caution due to a lack of power resulting from the exclusion of the second block for all participants. It is likely that the chance-level structural effects in the classification condition were a result of this lack of power. This finding will not be discussed further at this time as it did not occur in the next experiment. With that in mind, Experiment 4 was successful in eliminating the ceiling effect found for old endorsements in Experiment 3. Without the ceiling effect, old endorsements increased with study strength whilst NC and NI endorsements stayed the same. The likely scenario is of an increase in the strength of evidence of the old distribution, but without a criterion shift or a change in the strength of evidence for the NC and NI distributions.

The fact that criterion placement again appeared to be unaffected by the study-strength manipulation is unusual. The task is relatively complex with three different decisions being required for each trial. Task complexity can prevent criterion movement (Hockley & Niewiadomski, 2007). This was unlikely to be the case here as the combined analysis of Experiments 3 and 4 demonstrated that participants moved at least their recognition decision criterion in response to learning conditions with similar tasks at test. Also, participants may not have realised that study strength could be used as a performance aid, in which case the participants in the strong-study condition would have no reason to use a different criterion than those in the weak-study condition. Stretch and Wixted (1998) obtained a between-list study-strength mirror effect, although all participants took part in both weak-study and strong-study conditions and so the distinction between the two was obvious to each individual. Verde and Rotello's (2007) results suggested that participants do take notice of strength of the targets at test but only in the first block. If so, participants should be shifting their criterion by study strength in Experiment 4 because they would adopt

a more conservative criterion in the strong-study condition compared to the weak-study condition. It is reasonable to suspect that participants are not aware that the study strength can be used to set their criterion as each participant is only exposed to one strength condition. Experiment 5 informed participants of the fact that they will see words either once or multiple times in order to attempt to induce them to use this information in setting their criterion.

The differences between the recognition and classification task were also consistent with Experiment 3. Experiment 4 was not aimed at providing additional evidence to explain these patterns, and so the same possibilities exist as in Experiment 3. The old distribution and the criterion may be lower on the strength-of-evidence scale in classification compared to recognition, or else both the NC and NI distributions are higher.

Performance in the attribution categories indicated that there was no effect of study strength in the intuition attributions. Only memory attributions to old stimuli were responsive to study strength. This could represent recollection-based processes being attributed to memory which are known to be sensitive to study-strength manipulations (Yonelinas, 2002). Intuition responses to old words were not sensitive to study strength, but if intuition responses represented familiarity then they should increase by study strength (Jacoby, 1999; Yonelinas, 2002). It is possible that the study-strength manipulation is not affecting familiarity at all in this experiment. The next experiment will employ a stronger manipulation of strength to test this possibility. Another possibility is that intuition attributions do not represent familiarity as represented in the recognition literature, and that memory attributions include some aspect of familiarity. This possibility will be discussed more fully after Experiment 5.

Overall, Experiment 4 was consistent with the main findings from Experiment 3. There was a differential effect of study strength on the episodic and structural *d'* but participants did not shift their criterion from weak-study to strong-study conditions. Experiment 5 modified the design so that the study strength is more obvious so participants can use this information in setting their criterion. Additionally, the strong-study condition involved more repetitions than in Experiment 4 in order to see if the structural effect was not increasing because of a poor study-strength manipulation.

3.4 Experiment 5

3.4.1 Method.

3.4.1.1 Participants.

Sixty-four undergraduates from the UoS participated in the experiment. All participants were either given course credits or £5 payment.

3.4.1.2 Materials and design.

The materials and counterbalancing were the same as in Experiment 4, except for a small number of stimuli (two in total) that were replaced because they contained fragments of other stimuli such as earth and earthworm. The general design was the same but for the following changes. To try to make the study-strength manipulation more noticeable, participants were warned how many times they would see each word in the study phase. Additionally, words in the strong-study condition were displayed five times each instead of three. A retention interval was added between the study and test phases such that the study phase lasted ten minutes for every participant. During the interval, the participant solved simple maths problems. Task was changed to a between-subject variable such that only one test phase was administered to each participant with half of the participants being given a recognition test phase and the other half a classification test phase. Test lists were counterbalanced between participants such that each of the four test lists were used equally. When the questionnaire was administered, participants answered either Question 1a or 1b depending on if they had completed recognition or classification (see Appendix B).

3.4.1.3 Procedure.

The procedure was identical to Experiment 4, except for two changes. Each participant was administered only one test phase which was a recognition phase for half the participants and a classification task for the other half. A retention interval was used in order to equate the time from the beginning of the study phase to the test phase at ten minutes for all participants. In the retention interval participants had to solve maths problems in which a sum was presented with some digits missing (i.e. 45 + ?? = 59). Participants were asked to write down the missing digits.

3.4.2 Results.

Endorsement rates were calculated and corrected as in Experiment 4. One participant was excluded from the analysis for low recognition performance (more than three standard deviations below the mean).

3.4.2.1 Analysis of episodic and structural d'.

The episodic and structural *d'* measures were again compared to chance using confidence intervals. As in Experiment 4, the changes in magnitude of the *d'* measures were investigated with a 2 x 2 x 2 ANOVA with task type, effect type and study strength. See Table 3.16 for *d'* means and standard errors and Table 3.17 for the results of the ANOVA and associated pairwise statistics.

Table 3.16

d' by Study Strength, Task Type and Effect Type from Experiment 5 (SE in brackets)

Task and *d'* type	Weak-study	Strong-study	Total
Recognition			
Episodic *d'*	.90 (.09)*	1.89 (.22)*	1.38 (.15)
Structural *d'*	.32 (.10)*	.29 (.10)*	.30 (.07)
Classification			
Episodic *d'*	.58 (.13)*	1.80 (.23)*	1.19 (.17)
Structural *d'*	.34 (.08)*	.33 (.10)*	.33 (.06)
Both tasks			
Episodic *d'*	.74 (.08)	1.85 (.16)	1.29 (.11)
Structural *d'*	.33 (.06)	.31 (.07)	.32 (.05)
Total	.53 (.07)	1.08 (.07)	.81 (.04)

* = Lower bound of 95% confidence interval above chance level of 0.

Table 3.17

Results of 2 (Study Strength) x 2 (Task Type) x 2 (Effect Type) ANOVA on d' from Experiment 5

Effect	*F* value (df)	*p* value	η^2
Effect type	$F(1, 59) = 73.66$	.001*	.55
Study strength	$F(1, 59) = 38.54$	< .001*	.85
Effect type by study strength	$F(1, 59) = 24.11$	< .001*	.29
Study strength for episodic d'	$F(1, 59) = 38.45$	< .001*	.39
Study strength for structural d'	$F < 1$	-	-
There were no other significant effects	$F(1, 59) = 1.07$	.30	.02

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

The effect-type main effect reflected a higher episodic *d'* than structural *d'*. The study-strength main effect reflected better discrimination in the strong-study condition than the weak-study condition.

The interaction was further investigated with pairwise comparisons. As in Experiment 4, the episodic effect was greater in strong-study rather than weak-study conditions, whilst the structural *d'* did not change from weak-study to strong-study conditions (See Figure 3.5 and Table 3.16).

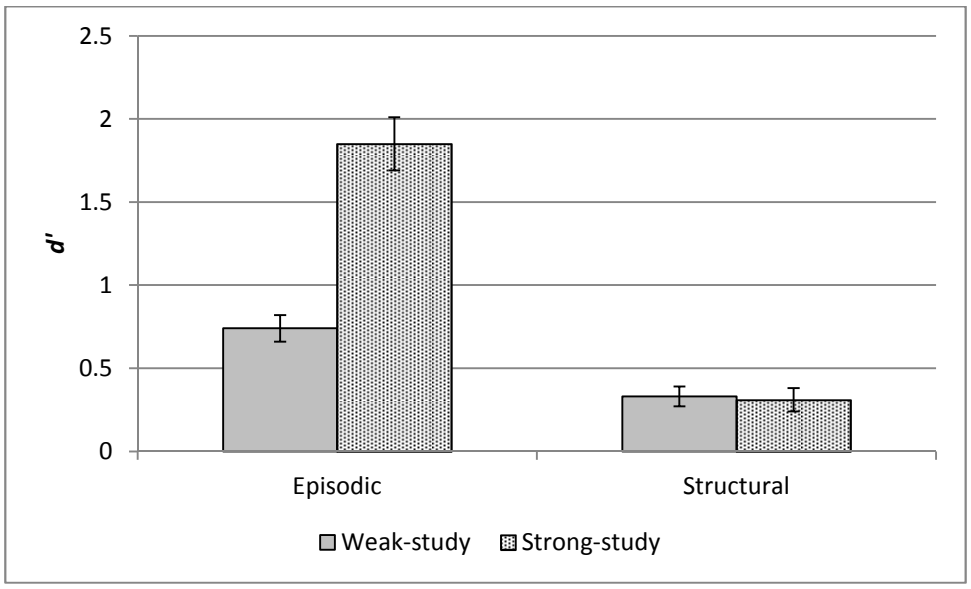

Figure 3.5. Episodic and structural effects by study strength in Experiment 5.

3.4.2.2 Analysis of endorsement rates.

Endorsement rates were again compared across study strength, word type and task types with an ANOVA. See Table 3.18 for means and standard errors and Table 3.19 for the ANOVA results and associated pairwise statistics.

Table 3.18

Endorsement Rates by Study Strength, Word Type and Task Type from Experiment 5 (SE in brackets)

Task and word type	Weak study	Strong study	Total
Recognition			
Old	.67 (.03)	.81 (.03)	.74 (.02)
NC	.34 (.03)	.22 (.03)	.28 (.03)
NI	.24 (.03)	.15 (.03)	.20 (.03)
Classification			
Old	.70 (.03)	.86 (.04)	.78 (.02)
NC	.50 (.05)	.36 (.05)	.43 (.03)
NI	.38 (.05)	.25 (.04)	.32 (.03)
Both Tasks			
Old	.69 (.02)	.84 (.02)	.76 (.02)
NC	.42 (.03)	.29 (.03)	.35 (.02)
NI	.31 (.03)	.20 (.03)	.26 (.02)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 3.19

Results of 3 (Word Type) x 2 (Task Type) x 2 (Study Strength) ANOVA on Endorsement Rates from Experiment 5

Effect	*F* value (df)	*p* value	η^2
Task type	$F(1, 59) = 11.72$	< .001*	.17
Word type	$F(2, 118) = 329.66$	< .001*	.85
Old versus NC	$F(1, 63) = 168.83$	< .001*	.73
NC versus NI	$F(1, 63) = 44.44$	< .001*	.41
Word type by study strength	$F(2, 118) = 27.19$	< .001*	.31
Study strength for old words	$F(1, 59) = 19.35$	< .001*	.25
Study strength for NC words	$F(1, 59) = 8.03$	.006*	.12
Study strength for NI words	$F(1, 59) = 8.08$	.006 *	.12
Word type and task type	$F(2, 118) = 3.68$	.03*	.06
Study strength for old words	$F(1, 59) = 1.68$	.20	-
Study strength for NC words	$F(1, 59) = 12.04$	.001*	.17
Study strength for NI words	$F(1, 59) = 9.51$	.003*	.14
There were no other significant effects	$F < 1$	-	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.
* = *p* value denotes significant difference.

The word-type main effect reflected more old endorsements ($M = .76$, $SE = .02$) than NC endorsements ($M = .35$, $SE = .02$) and more NC endorsements than NI endorsements ($M = .26$, $SE = .02$). The task-type main effect reflected more endorsements overall in classification ($M = .51$, $SE = .02$) than in recognition ($M = .41$, $SE = .02$).

The word-type by study-strength interaction was due to an increase in old endorsements from weak-study to strong-study conditions and a decrease in NC and NI endorsements from weak-study to strong-study conditions. The word-type and task-type interaction was due to NC and NI endorsements increasing from recognition to classification whilst old endorsements did not change.

This pattern of results is consistent with the initially expected pattern of an increase in old endorsements, a decrease in NC and NI endorsements and a criterion shift – Case 1 in Table 3.1.

3.4.2.3 Performance by attribution.

Endorsement rates were again broken down into attribution types and entered into a 3 x 2 x 2 repeated measures ANOVA with word-type (old, NC and NI), study strength (weak study versus strong study) and task type (recognition versus classification). Guess and rules were again excluded. See Table 3.20 for means and standard errors and Table 3.21 for the results of the ANOVAs and associated pairwise statistics.

Table 3.20

Mean Endorsement Rates by Attribution, Word Type, Task Type and Study Strength from Experiment 5(SE in brackets)

Task and word type	Weak study		Strong study	
	Intuition	Memory	Intuition	Memory
Recognition				
Old	.23 (.03)	.34 (.03)	.10 (.02)	.62 (.04)
NC	.17 (.02)	.06 (.01)	.09 (.03)	.08 (.02)
NI	.11 (.02)	.06 (.01)	.08 (.02)	.04 (.01)
Classification				
Old	.14 (.02)	.37 (.04)	.15 (.04)	.63 (.07)
NC	.15 (.02)	.12 (.03)	.16 (.03)	.05 (.01)
NI	.12 (.02)	.08 (.02)	.13 (.03)	.03 (.01)
Both tasks				
Old	.18 (.02)	.35 (.04)	.12 (.02)	.63 (.04)
NC	.16 (.02)	.09 (.01)	.12 (.02)	.06 (.01)
NI	.12 (.02)	.07 (.01)	.10 (.02)	.03 (.01)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 3.21

Results of 3 (Word Type) x 2 (Task Type) x 2 (Study Strength) ANOVAs on Intuition and Memory Endorsement Rates from Experiment 5

Effect	*F* value (df)	*p* value	η^2
Intuition ANOVA			
Word type	$F(2, 118) = 6.91$	.001*	.10
Old versus NI	$F(1, 63) = 9.20$	.004*	.13
NC versus NI	$F(1, 63) = 14.28$	.001*	.18
Study strength by task type	$F(1, 59) = 4.20$	.04*	.07
Study strength for recognition	$F(1, 59) = 7.27$	.009*	.11
Study strength for classification	$F < 1$		
There were no other significant effects	$F(1, 59) = 3.22$	.08	-
Memory ANOVA			
Word type	$F(2, 118) = 263.40$	< .001*	.82
Old versus NC	$F(1, 63) = 166.01$	.001*	.72
NC versus NI	$F(1, 63) = 11.64$	.001*	.16
Study strength	$F(1, 59) = 11.19$	.001*	.16
Word type by study strength	$F(2, 118) = 33.67$	.001*	.36
Study strength for old words	$F(1, 59) = 28.77$	.001*	.33
Study strength for NC words	$F(1, 59) = 2.45$	.12	-
Study strength for NI words	$F(1, 59) = 4.70$	.03*	.07
There were no other significant effects	$F(1, 59) = 1.58$	.21	-

Note - Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.
* = *p* value denotes significant difference.

In intuition, the word-type main effect reflected fewer intuition endorsements for NI words ($M = .11$, $SE = .01$) than for old ($M = .15$, $SE = .01$) or NC words ($M = .14$, $SE = .01$). The study-strength by task-type interaction reflected the fact that in recognition there were fewer intuition endorsements in the strong-study ($M = .09$, $SE = .02$) than the weak-study condition ($M = .17$, $SE = .02$), whilst in classification there was no change from weak-study ($M = .14$, $SE = .02$) to strong-study ($M = .14$, $SE = .02$) conditions.

In memory, the study-strength main effect reflected more memory endorsements in the strong-study condition ($M = .24$, $SE = .01$) than the weak-study condition ($M = .17$, $SE = .01$). The word-type main effect reflected more memory endorsements for old words ($M =$

.49, *SE* = .02) than for NC (*M* = .08, *SE* = .01) and more for NC than for NI (*M* = .05, *SE* = .01). The interaction represented an increase in memory endorsements for old words from weak-study to strong-study conditions and a decrease in NI endorsements from weak-study to strong-study conditions. There was no change in NC endorsements by study-strength.

3.4.2.4 Awareness questionnaire.

Once again, the questionnaire indicated that no participants developed explicit knowledge of the rule set.

3.4.3 Discussion.

Experiment 5 once again found that the episodic effect was sensitive to study strength whilst the structural effect was not. This pattern was obtained despite the enhancement of the episodic effect by the increase in the number of repetitions in the strong-study condition (five versus three). This suggests that a poor strength manipulation was not responsible for the structural effect not increasing with study strength in the previous experiments. The main differences between Experiments 5 and 4 were that participants were informed about the study-strength manipulation and that the study-strength manipulation was slightly more extreme. This had the desired effect – participants in Experiment 5 have shifted their criterion consistent with Case 1 in Table 3.1. The old distribution shifted to the right from weak-study to strong-study conditions, as did the criterion. Consistent with previous experiments, there were more NC and NI endorsements in classification than in recognition. It is unlikely that the NI distribution would move between these conditions, so the most likely explanation is that participants adopted a more liberal criterion for the classification task and paid less attention to the old stimuli in making their decision. It has been shown before that task demands can affect how knowledge is used (Whittlesea & Dorken, 1993). Kinder, Shanks, Cock and Tunney (2003) demonstrated that including old items at test induced participants to use recollection processes for classification type decisions. In the experiments presented here only a third of test items were old, so perhaps this low proportion resulted in participants failing to use recollection processes as much as they should in the classification task. The fact that memory attributions were used equally in recognition and classification argues against a lack of recollection in classification. Experiments 7 and 8 will attempt to provide further clarity on this issue by reducing recollection to see if the same pattern holds even when recollection is already impaired.

The attribution data suggested that participants used their knowledge differently in each task, or at least that participants had different subjective beliefs about what knowledge they were using. An increase in study strength resulted in fewer intuition endorsements across the board for recognition but not for classification. In recognition, participants may have abandoned the use of intuition-type feelings when they perceived that the study list was memorable due a high number of repetitions of the words at study. This is similar to the change in tactics seen with differing global subjective memorability (Bruno et al., 2009; Higham et al., 2010). In classification, participants may have been aware that they did not know the rule set, and so were more likely to rely on intuition regardless of how memorable the study list seemed to them. The memory endorsement data are similar to the previous experiment, and could represent the influence of recollection or familiarity or both.

The most relevant question posed by these results is why the structural effect was not sensitive to study strength whilst the episodic effect was. Most recognition theories suggest that recognition performance is based on two processes – recollection and familiarity. One example of such theory is the source of activation confusion (SAC) model (Reder et al., 2000). The SAC model is a network-type model in which nodes are created for individual words. These nodes are connected to further nodes that represent various characteristics about the words. Activation of any node strengthens itself as well as any nodes to which it is connected. Decisions are then based on the strength of activation of these nodes. An initial recognition decision is made on the basis of a recollection attempt using study-context nodes. If this results in no recollection, then the general familiarity associated with the word node is used – if the familiarity is high, then false endorsements are made. This model predicts that recollection-based performance should increase with study strength as nodes representing the study context would have a higher strength of evidence in strong-study rather than weak-study conditions. Familiarity based endorsements would also increase with study strength as the nodes associated with rule consistence would be activated more in the strong-study condition than the weak-study condition and this activation would increase the strength of all possible rule-consistent words relative to rule-inconsistent words. However, models such as this are insufficient to explain performance in the recognition task for Experiments 3, 4 and 5 – if familiarity were affected by study strength then the structural effect should have been larger in strong-study rather than weak-study conditions.

Implicit-learning dual-process theories suggest a process based on explicit structural information and one based on implicit structural information (e.g. Scott & Dienes, 2008). Performance based on implicit structural knowledge is reflected in the CFM (see section

3.2.3), whilst performance based on explicit structural knowledge reflects consciously held rules-based knowledge in addition to implicit familiarity. Performance without awareness is based on small amounts of familiarity that participants use for their judgements, but do not trust enough to reflect in their confidence ratings. However, this is also insufficient to explain the results. By this model, there should again be a larger structural effect in strong-study rather than weak-study conditions, as greater amounts of familiarity would be associated with rule-consistent words than rule-inconsistent words. In fact, the structural effect stays constant across both task type and study strength. See Chapter 5 for more on specific models and their predictions for the structural effect.

In order to explain the results, two types of familiarity are required. One type of familiarity, "structural familiarity", drives the discrimination between NC and NI words. The second type of familiarity, "episodic familiarity", is the type of familiarity usually cited in recognition. Episodic familiarity combines with an explicit episodic recollection process to produce the discrimination between NC and old words. The results found can now be explained – episodic familiarity (and also episodic recollection) is affected by the study-strength manipulation whilst structural familiarity is not. The question of what exactly underlies this structural familiarity is difficult, as most theories of both recognition memory and implicit learning would predict some kind of sensitivity (and thus movement on the NC distribution) by study strength. Factors that might prevent structural familiarity being resistant to study strength are discussed in Chapter 6.

A two-familiarity theory poses a problem for the attribution data. Although NC and NI intuition and memory responses could both reflect structural familiarity, it is hard to pinpoint where episodic familiarity might be placed in the attributions. It is likely to depend on participants' metacognitive beliefs. If a participant often uses familiarity as a memory aid, then they may use the memory attribution for familiar responses. On the other hand, they may interpret a feeling of familiarity as intuition. In order to investigate the dual-familiarity theory, a different set of attributions may be needed.

In Experiments 3, 4 and 5, I have started to examine a paradigm that is similar to both recognition experiments and to implicit learning experiments. The invariance of the structural effect to study strength suggests an answer to the question posed at the end of Chapter 2 concerning how similar recognition memory familiarity is to implicit learning familiarity. Since recognition literature familiarity increases with study-strength (Jacoby, 1999), and performance based on learning rule sets is based on familiarity (Scott & Dienes, 2008), familiarity as cited in the recognition literature is not quite the same thing as

familiarity as used in the implicit learning literature. This is not to imply that participants can tell the difference between each type of familiarity – in fact participants' attitudes to their own knowledge seem to depend on what task they are involved in, even though the actual patterns of performance are similar. The basis of the familiarity certainly seems different – the structural effect is driven by some sort of knowledge of the structure whilst episodic familiarity is driven by memory of having seen the word before. The structural effect is insensitive to study strength, whilst episodic familiarity increases by study-strength.

There are several alternative explanations for the finding that the structural effect is insensitive to study strength. The structural effect may be a small and noisy effect, a limitation which Experiment 6 attempts to address. Episodic familiarity could be reflected in the structural effect and not the episodic effect. Chapter 4 attempts to address this question. Participants could be using simple surface characteristics to make their decisions, and some recognition-memory models might be able to replicate the data. These possibilities are addressed in Chapter 5.

3.5 Experiment 6

3.5.1 Introduction.

One possible criticism of the claim that the structural effect is not sensitive to study strength is that it may be a small and noisy effect. This could mask changes in the structural effect by study strength. Experiment 6 will therefore use several different manipulations to increase the magnitude of the structural effect. It is possible that participants may learn the rule set better if they see more words at once, such that relations can be drawn between them. Thus in two of Experiment 6's conditions, 20 words will be displayed on the screen at once. Also, although direct instructions to learn the rule set would probably fail, it may be possible to indirectly increase the efficiency of participants' rule learning by asking them to look for commonalities between the study words. Finally, it is likely that the structural effect is familiarity-based like other rule-learning effects (e.g. Scott & Dienes, 2008). Both familiarity and recollection are thought to be sensitive to display time at study (Yonelinas, 2002). If this is the case, then reducing the display time should result in a change in the magnitude of the structural effect. Thus the time that words were displayed at study was reduced.

A second criticism is that the signal-detection measure *d'* might not be appropriate for the structural effect because the variance of the NC and NI distributions may differ. It is thought that the variance of the old and new distributions differ (Wixted, 2007) with the old

distribution having a larger variance than the new distribution. The *d'* measure is less appropriate in these circumstances although it seems resilient enough to still find popular use for measuring old/new discrimination. Thus Experiment 6 will also look at the variance of the NC and NI distributions.

3.5.2 Method.

3.5.2.1 Participants.

Forty-nine undergraduates from the UoS participated in the experiment. All participants were either given course credits or £5 payment.

3.5.2.2 Materials.

The materials were the same as in Experiment 5.

3.5.2.3 Design.

The design of the experiment was similar to Experiment 5 except study strength was not manipulated. Instead, the manner of display of the study words was manipulated to create three conditions – "fast-display" in which stimuli were presented rapidly; "20-words-recogniton" in which words were presented 20 at a time with recognition instructions and "20-words-commonality" in which words were presented 20 at a time with instructions to look for commonalities between them. Due to budget constraints there was only a recognition task. Thus the design consisted of a between-subjects manipulation of the study task. The weak-study condition of Experiment 5 was used as a baseline condition with which to compare the three study-task conditions.

3.5.2.4 Procedure.

The procedure was the same as Experiment 5 other than for the following changes. Participants were split equally between three study conditions. Participants in the fast-display condition were presented words individually and told to remember them for a later recognition test. The words were displayed for 250 ms with a 250 ms inter-stimulus-interval (ISI). Participants in the 20-words-recognition condition were presented the words in sets of 20 at a time. Each set was on screen for 20 seconds and participants were asked to remember the words for a later memory test (i.e. the 20 words remained on screen for 20 seconds and were then replaced with the next 20 until all 80 words in the study-list had been presented). Participants in the 20-words-commonality condition were also presented the study words in

sets of 20 but were instructed to look for commonalities between the words and were not told in advance about the recognition test.

At test, all participants were given recognition instructions. Participants were asked to identify words they had seen in the study phase using a 6 point scale (1 = sure new; 2 = fairly sure new; 3 = guess new; 4 = guess old; 5 = fairly sure old; 6 = sure old). No other rating or phenomenological rating was required. This rating scale was used in order to make the analysis of distribution variance possible. The experiment ended with the same awareness questionnaire as was used in Experiment 5.

3.5.3 Results.

One participant was excluded from the analysis as they correctly selected the individual elements of the conjunctive rule-set on the questionnaire, although they did not correctly identify the nature of the conjunction

3.5.3.1 Analysis of episodic and structural d'.

The *d'* measures from this experiment were compared with the weak-study *d'* measures from Experiment 5. Herein the weak-study condition from Experiment 5 is referred to as the baseline study-task. Study task and effect type were entered into a 4 x 2 ANOVA with a between-subject factor of study task (20-words-recognition versus 20-words-commonality versus fast-display versus baseline) and a within-subject factor of effect type (episodic versus structural). See Table 3.22 for means and standard errors.

Table 3.22

d' by Effect Type and Study Task from Experiment 6 (SE in Brackets)

d' type	Baseline[1]	20-words-recognition	20-words-commonality	Fast-display
Episodic *d'*	.90 (.09)*	.92 (.14)*	.68 (.09)*	.89 (.13)*
Structural *d'*	.32 (.10)*	.23 (.05)*	.17 (.09)	.19 (.06)*

1 = study 5 weak-study means and SEs.

* = Lower bound of 95% confidence interval above chance level of 0.

Only the main effect of effect-type was significant, $F(1, 60) = 74.99$, $p < .001$, $\eta^2 = .56$. This reflected the fact that the episodic *d'* ($M = .85$, $SE = .06$) was higher than the

structural *d'* (M = .23, SE = .04) in all conditions. No other effects were significant, highest $F(1, 60) = 1.37$, $p = .26$.

3.5.3.2 Analysis of endorsement rates.

Endorsement rates were entered into a 4 x 3 ANOVA with a between-subject factor of study task (20-words-recognition versus 20-words-commonality versus fast-display versus base) and a within-subject factor of word type (old versus NC versus NI).

Table 3.23

Endorsement Rates by Study Task and Word Type (SE in Brackets)

Task and word type	Baseline[1]	20-words-recognition	20-words-commonality	Fast-display
Old	.67 (.03)	.68 (.03)	.70 (.02)	.70 (.03)
NC	.34 (.03)	.35 (.04)	.45 (.03)	.37 (.03)
NI	.24 (.03)	.28 (.04)	.38 (.05)	.32 (.04)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.
1 = Experiment 5 weak-study means.

Only the word-type main effect was significant, $F(2, 122) = 239.63$, $p < .001$, $\eta^2 = .80$. This reflected more old endorsements overall (M = .68, SE = .01) than NC endorsements (M = .37, SE = .02), $F(1, 64) = 254.09$, $p < .001$, $\eta^2 = .80$ and more NC endorsements than NI endorsements (M = .30, SE = .02), $F(1, 64) = 29.07$, $p < .001$, $\eta^2 = .31$. There were no other significant effects, highest $F(1, 61) = 2.10$, $p = .11$.

3.5.3.3 PLUM Analysis.

A PLUM analysis was carried out, which uses participants' responses on the rating scale to different stimuli types to estimate whether the underlying distributions are of equal or unequal variance. The analysis attempts to match two models to the data – one where the distributions being looked at have equal variance and one where they have unequal variance. A non-significant result on a χ^2 goodness-of-fit test indicates that the model under inspection is a good fit to the data. For more details see Decarlo (2003). An analysis was carried out for each of the three study tasks (excluding the baseline) and separate analyses were carried out to compare the old and NC distributions and the NC and NI distributions.

For old and NC stimuli, an unequal-variance model was the best fit with 20-words-recognition $\chi^2(3, N = 1280) = 5.11$, $p = .16$, 20-words-commonality $\chi^2(3, N = 1280) = 5.04$, $p = .17$ and under fast-display conditions $\chi^2(3, N = 1280) = 4.87$, $p = .18$. On the other hand, the analysis for the NC and NI stimuli showed that the best fit was an equal-variance model with 20-words-recognition $\chi^2(4, N = 1280) = .60$, $p = .96$, 20-words-commonality $\chi^2(4, N = 1280) = 2.14$, $p = .71$ and under fast-display conditions $\chi^2(4, N = 1280) = 3.04$, $p = .55$. To summarise, the old and NC distributions did not have equal variance whilst the NC and NI distributions did have equal variance. PLUM provides a parameter (*a*) that can be used to estimate the ratio of the standard deviations of the distributions, by taking the exponential of *–a*. Doing so yielded NC to old ratios of 0.79, 0.82 and 0.87 for the 20-words-recognition, 20-words-commonality and fast-display conditions respectively.

3.5.4 Discussion.

None of the steps taken to change the structural effect seem to have worked. Number of words seen at once, display time and instructions all resulted in no change in the structural effect from that observed in Experiment 5. Instructions to look for commonalities in the words may have decreased the structural effect a little, and perhaps the episodic effect too, but not to a statistically significant extent. The structural effect appears to be quite resistant to change. This is discussed further in Chapter 6.

There were no changes in the endorsement rates in any condition. It is surprising that participants did not at least respond to the difficult study conditions in fast-display by adopting a more liberal criterion related to Experiment 5. Numerically, the NC and NI endorsement rates are in the right direction for this to be the case. Overall, the endorsement rate analysis supported the fact that the changes in study task did not alter the magnitude of the structural effect.

The PLUM analysis validated the use of the *d'* measure, at least for measuring the structural effect. What model best describes the old and new distributions in signal-detection models has received a lot of attention in the literature (Mickes, Johnson, & Wixted, 2010; Onyper, Zhang, & Howard, 2010; Wixted, 2007) but possible differences between different types of new distribution have received little attention. The PLUM analysis indicated that the NC and NI distributions were of equal variance and thus the signal-detection measure *d'* is appropriate.

The PLUM analysis indicating that the old and NC distributions did not have equal variance is consistent with previous research (Wixted, 2007). The NC to old standard deviation ratios of 0.79, 0.82 and 0.87 are consistent with previous results which yielded lure to old ratios of 0.80, suggesting that the lure distributions have a smaller standard deviations than old distributions (e.g. Mickes, Wixted, & Wais, 2007). With unequal variance distributions a different measure is recommended called d_a (Macmillan & Creelman, 2005). There are several reasons why *d'* will continue to be the main measure used in the experiments presented here. As Mickes et al. (2007) point out, the d_a measure is not often used because it requires an ROC analysis. Adding ROC analyses would complicate and lengthen the results. If the episodic effect were the central focus this may be worth the additional complexity but the focus of the experiments presented here is on the structural effect and as already noted *d'* is appropriate for the structural effect. The main effect of using *d'* with unequal variance distributions would be to either under- or over-estimate the actual discrimination. However, the actual value of the *d'* is not of interest in most experiments here, it is whether or not the relevant *d'* is sensitive to the study-strength manipulation. Even if the episodic *d'* is over- or under-estimated, the direction of change of the episodic *d'* with respect to the study-strength manipulation would not be affected by the use of *d'* rather than d_a. Additionally, the observed changes in the episodic *d'* from Experiments 3, 4 and 5 were consistent with changes in episodic processes such as recollection and familiarity (Yonelinas, 2002). Finally, analysis of the pure endorsement rates were consistent with changes in the episodic *d'*, further suggesting that *d'* is a sufficiently acceptable measure of changes due to study strength.

All of the results in Chapter 3 point to the structural effect being insensitive to study strength. In Chapter 4, two experiments are presented that take a closer look at whether episodic familiarity and structural familiarity are in fact the same.

.

This page intentionally left blank

4 Chapter 4: Reducing Recollection

4.1 Introduction

Chapter 3 advanced the theory that structural familiarity and episodic familiarity are different with structural familiarity representing the influence of rule-based knowledge and episodic familiarity representing memory for the stimuli. If this is the case then the episodic effect is made up of both recollection and episodic familiarity whilst the structural effect represents structural familiarity. An alternative explanation is that structural familiarity is not different from episodic familiarity and the influence of episodic familiarity is responsible for the structural effect in some way. In order to determine whether different knowledge types underpin episodic and structural familiarity, it is necessary to eliminate the influence of recollection. This should leave only familiarity at work in producing both the episodic effect and structural effect and allow a direct comparison of the influences driving each effect. With recollection reduced, if the episodic effect still increases with study strength then this increase will be due to episodic familiarity. If the structural effect remains invariant to study strength, then episodic familiarity cannot drive the structural effect, because episodic familiarity is sensitive to study strength (Jacoby, 1999; Yonelinas, 2002). On the other hand, if reducing recollection results in the episodic effect becoming invariant to study strength along with the structural effect, then this would suggest that both types of effect are underpinned by episodic familiarity. Experiment 7 attempted to reduce the influence of recollection whilst leaving episodic familiarity intact.

Imposing a deadline is known to reduce recollection whilst leaving familiarity intact (Jacoby, 1999; Yonelinas, 2002). Jacoby asked participants to read a list of words, either once, twice or three times. Then participants heard another set of words. At test, participants were asked to say yes to words that they had heard before but no to words they had seen before. Deadlines were imposed at test in order to reduce the effect of recollection. The idea was that participants could reject seen words by recollecting the experience of seeing them. However, if the influence of that recollection was reduced then participants would have a strong feeling of familiarity for words they had read three times, but be unable to reject the word because they could not recollect the source of the word. The results were consistent with this – participants asked to make the decision with a short deadline made more FAs to words read three times than to those read only once whereas participants asked to make the decision with a long deadline made fewer FAs to words read three times than to those read once. Experiment 7 uses a similar design in order to reduce the effects of recollection.

From a signal-detection point of view, the strength of evidence for a word can be thought of as being made up of contributions from familiarity and recollection in different proportions, both of which could operate as their own signal-detection models (Wixted & Mickes, 2010). If the structural effect is small and noisy, then previous experiments may not have detected changes due to study strength. If this is the case and the structural effect is based on familiarity, then increasing the proportion of familiarity that contributes to the final strength of evidence would reduce the noise and make changes easier to detect. Familiarity is thought to be faster than recollection (Yonelinas, 2002) and so imposing a deadline should increase the relative contribution of familiarity to the final strength of evidence compared to recollection.

The exact deadlines used will be those used by Jacoby (1999). A long-deadline condition barred participants from responding until 1.25 seconds from the appearance of the stimuli had passed. Participants then had 0.75 seconds to input their response. The short deadline gave participants just 0.75 seconds from the appearance of a stimulus to respond. In other words with a long deadline participants had up to two seconds to consider and give their response (1.25 seconds thinking time and 0.75 seconds response time) whilst with a short deadline participants had only 0.75 seconds to think and give a response.

4.2 Experiment 7

4.2.1 Predictions.

As in Chapter 3, the predictions are set in a signal-detection framework by discussing the possible effects of the manipulations on the old, NC and NI strength-of-evidence distributions. In the long-deadline condition, recollection will be reduced relative to previous experiments but should still have an influence (as demonstrated by the long deadline in Experiment 1 of Jacoby, 1999). Thus in the long-deadline condition the study-strength manipulation will have a similar effect on NC and NI distributions as it did in Experiment 5. The increase in study strength will shift the criterion to the right and increase the strength of evidence of the old distribution, leading to a reduction in NC and NI endorsements and an increase in old endorsements. As in Experiment 5, the episodic effect will be sensitive to study strength whilst the structural effect will be insensitive to study strength.

Imposing a short deadline will result in participants judging the task more difficult compared to the long-deadline condition. To compensate, participants will use a more liberal criterion in the short-deadline condition compared to the long-deadline condition (the shift in

C from Panel A to Panel B in *Figure 4.1*). The position of the NC and NI distributions should not be affected by the deadline manipulation. Thus the criterion shift will reduce NC and NI endorsements from long to short deadlines. Recollection will be reduced with a short deadline compared to a long deadline. This will result in the old distribution shifting to the left from long to short deadlines (the change in the old distribution from Panel A to Panel B in *Figure 4.1*). It is difficult to say how the old endorsement rate will change, as it depends on the extent to which the deadline manipulation reduces recollection. An increase in old endorsements would suggest that the change from long to short deadline shifted the criterion to the left to a greater extent than the shift in the old distribution. No change would suggest that the old distribution and criterion shifted left to an equal extent from long to short deadline and a decrease in old endorsements would suggest that the old distribution shifted to the left more than the criterion from long to short deadline.

It is assumed that in the short-deadline condition recollection is greatly reduced (see section 4.2.4 for more on this assumption). Thus the changes by study strength in the short-deadline condition will depend on episodic familiarity. As before, the increase in study strength from weak to strong will cause participants to adopt a more conservative criterion. The exact changes in the endorsement rates will depend on what happens to the underlying distributions with study strength. The changes from Panel B to Panel C in *Figure 4.1* depict the case where the episodic effect is no longer sensitive to study strength, resulting in the old distribution staying static. In this case, the criterion shift from weak- to strong-study conditions will result in old, NC and NI endorsements all decreasing. As there is no distribution movement, both the episodic and structural effects stay static, suggesting episodic and structural familiarity are not different. The changes from Panel B to Panel D in *Figure 4.1* depict the case where the episodic effect is still sensitive to study strength. The NC and NI distributions will stay static and the old distribution will shift to the right from weak- to long-study conditions. Thus the increase in study strength will be reflected by a decrease in the NC and NI endorsement rates and either an increase or no change in old endorsement rates depending on the extent to which the old distribution shifts compared to the criterion. A decrease in old endorsements is also possible, but they will decrease less than NC and NI endorsements so as to support an increase in the episodic effect.

Finally, an increased reliance on familiarity could result in the structural effect becoming sensitive to study strength, because a previously noisy increase becomes easier to detect. If this occurs, then the NC distribution will shift to the right from weak- to strong-study conditions whilst the NI distribution stays static. Together with the criterion shift this

will result in the NI endorsement rate decreasing from weak- to strong-study conditions as before whilst the NC endorsement rate will increase or stay static resulting in an increase in the structural effect.

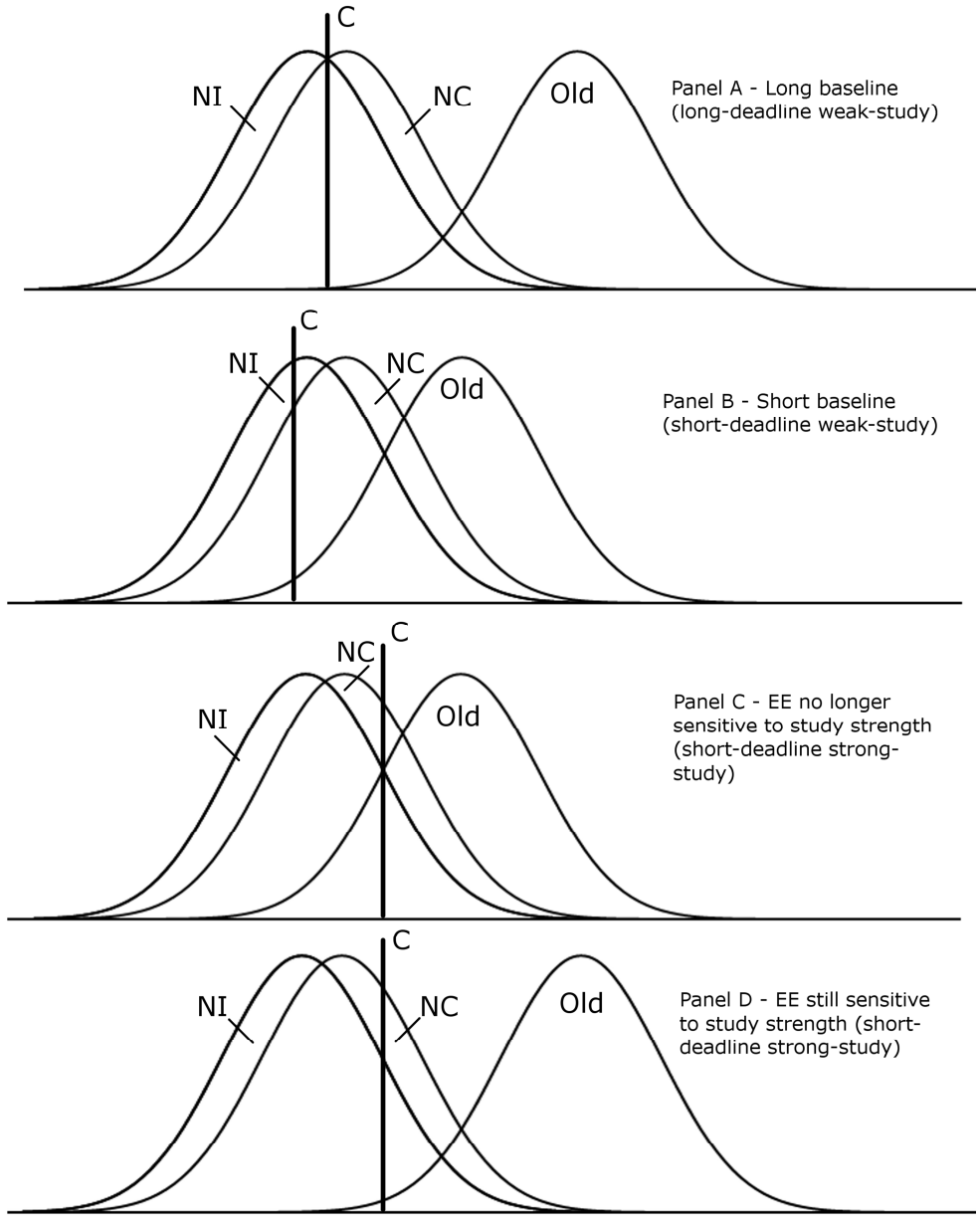

Figure 4.1. Patterns for changes in distributions and criterions due to study strength and deadline in Experiment 7. Old = old words; NC = new consistent words; NI = new inconsistent words, C = criterion, EE = episodic effect. Panels A and B depict the weak-study conditions with long and short deadlines respectively. Panels C and D depict the possible patterns of distributions for the strong-study condition with a short deadline. The diagrams depict a situation where the study-strength manipulation results in a large criterion shift and the deadline manipulation results in a small criterion shift so as not to confuse the two shifts. The actual criterion shift due to study strength could result in the criterion in Panels C and D being in the same place as that of Panel A.

4.2.2 Method.

4.2.2.1 Participants.

Sixty-four undergraduates from the UoS participated in the experiment. All participants were given either course credits or £5 payment for their time.

4.2.2.2 Materials.

The same materials as in Experiment 5 were used with the following modifications. The test lists were increased to 180 words in order to provide a larger word pool to split between short and long deadlines. Only Test Lists 1 and 3 were used. In order to increase the number of words on these test lists, 20 old words were taken from Test Lists 2 and 4 and added to Test Lists 1 and 3 respectively. Then 10 words of each category (CC, RA, CA, RA) were taken from Test List 2 and added to both Test Lists 1 and 3 such that NC words from one test list acted as NI words for the other test list and vice versa. Half of the words were assigned to a long deadline and half to a short deadline such that 30 old, 30 NC and 30 NI words were assigned to each deadline condition with each category of word equally represented in short and long deadlines. The words acting as short- and long-deadline words were counterbalanced, so that the short-deadline words for one participant acted as the long-deadline words for another and vice versa.

4.2.2.3 Design.

The design was similar to Experiment 5 with the following changes. A deadline manipulation was introduced. Each individual trial at test was either a short-deadline trial or a long-deadline trial. In the short-deadline trials a stimulus appeared on screen for a total of 0.75 seconds. Participants were required to respond to the stimulus before this time had elapsed. In the long-deadline condition the stimulus appeared for 1.25 seconds during which time participants could not respond. Once the 1.25 seconds had elapsed, asterisks appeared around the stimulus. This signified that the participant had 0.75 seconds to input a response. In other words, in the long-deadline condition participants were forced to view the stimulus for 1.25 seconds before being given 0.75 seconds to respond, whilst in the short-deadline condition participants had 0.75 seconds (total) to both view the stimulus *and* respond. All participants were given 90 long trials and 90 short trials, with 30 of each word type appearing in both short- and long-deadline conditions. Thus the overall design utilised a within-subject

deadline manipulation (short versus long), and between-subject manipulations of study strength (weak study versus strong study) and task type (recognition versus classification).

Due to pilot participants reporting fatigue from the total time of the experiment the requirement to provide a confidence rating was removed. In addition, instead of random chance, memory, intuition or rules the attribution choice was changed to "recollect", "familiarity" or "consistent" in an attempt to tap episodic and structural familiarity more directly. Participants were able to separately mark any of their responses as a guess with a radio button.

4.2.2.4 Procedure.

The study phase and retention interval were both identical to Experiment 5. After the retention interval, participants were given either recognition or classification instructions. They were also informed about the deadline manipulation and the phenomenological ratings. For each trial, four Xs appeared in the middle of the screen for one second. These were then replaced with either the word "fast" for short-deadline words or the word "slow" for long-deadline words. The word fast or slow stayed on screen for 1.5 seconds after which it disappeared. After a 0.5 second pause the target word then appeared. The participant was required to indicate if they believed the word was old or new in the recognition condition, or consistent or inconsistent with the rule set in the classification condition. Their choice was made by pressing either F or J on the keyboard. If the deadline was short the participant was given 0.75 seconds to respond to the word. If the deadline was long the word stayed on screen for 1.25 seconds during which the participant could not respond. A line of asterisks then appeared above and below the word, signifying that the participant now had 0.75 seconds to respond. If a participant tried to respond early, a beep sounded and the trial continued uninterrupted. If a participant did not respond by the deadline they were reminded that they had to respond quickly. In any case their actual response was only recorded if it was within the correct time period – late responses were simply recorded as a time-out trial and early responses were ignored. After responding, the participant was then invited to select the basis of their decision. There were different options for this choice depending on if they were in the recognition or classification condition. For recognition, if the participant responded old, a screen appeared where they could choose between recollect and familiarity as the basis of their decision using a radio button. If they chose new, they were asked if they believed the word was consistent or inconsistent with the rule set. On all response screens they could also indicate if they believed that their response was a guess. For classification, if

they selected consistent they were then asked if they thought the word was consistent because it was old or if they believed it to be a new word. If they selected old, they were then given the choice between recollect and familiarity as the basis of their decision. If they selected new then this was taken as a consistent attribution.

Once the participant had read the instructions, they were given eight trials in which to practice followed by a chance to ask questions before the test phase began. After the test phase, they were given a questionnaire as in Experiment 5.

4.2.3 Results.

Eight participants were excluded from the analysis. Four were excluded due to low cell counts through missing more than 50% of responses, two were excluded for non-compliance with the instructions and two participants were excluded because they correctly selected the individual elements of the conjunctive rule-set on the questionnaire, although they did not correctly identify the nature of the conjunction.

For the remaining participants, missed trials were discarded (a mean of 7% of long-deadline responses and 22% of short-deadline responses) and the endorsement rates were conditionalised on only those trials where a response was provided. In order to maintain an acceptable cell count, trials classified as guesses were analysed along with the other data and were not separated out. The attributions data analyses were complex and inconclusive and so are not presented here. Interested readers can refer to Appendix C.

4.2.3.1 Analysis of episodic and structural d'.

As in previous experiments, the *d'* were compared with chance performance with the lower bound of the 95% confidence interval. The results can be seen in Table 4.1. The changes in magnitude of the *d'* measures were investigated with a 2 x 2 x 2 x 2 ANOVA with between-subject factors of task type (classification versus recognition), study strength (weak study versus strong study), and within-subject factors of effect type (episodic versus structural), and deadline (short versus long). For all results and pairwise comparisons see Table 4.2.

Table 4.1

d' by Study Strength, Task Type, Effect Type and Deadline from Experiment 7 (SE in brackets)

Task and *d'* type	Short deadline		Long deadline		Total	
	Weak study	Strong study	Weak study	Strong study	Weak study	Strong study
Recognition						
Episodic *d'*	.32 (.16)*	1.26 (.19)*	.61 (.15)*	1.86 (.15)*	.46 (.13)*	1.56 (.15)*
Structural *d'*	.16 (.11)	.34 (.12)*	.30 (.14)*	.12 (.10)	.23 (.08)*	.23 (.09)*
Classification						
Episodic *d'*	.57 (.11)*	.83 (.20)*	.92 (.09)*	1.26 (.25)*	.75 (.13)*	1.04 (.14)*
Structural *d'*	.04 (.11)	.36 (.08)*	.19 (.11)	.26 (.15)*	.12 (.08)	.31 (.08)*

* = Lower bound of 95% confidence interval above chance level of 0.

Table 4.2

Results of 2 (Study Strength) x 2 (Task Type) x 2 (Effect Type) x 2 (Deadline) ANOVA on d' for Experiment 7

Effect	*F* value (df)	*p* value	η^2
Deadline	$F(1, 52) = 24.31$	< .001*	.32
Effect type	$F(1, 52) = 70.77$	< .001*	.58
Study strength	$F(1, 52) = 28.53$	< .001*	.35
Study strength by task type	$F(1, 52) = 4.17$	.05*	.07
Study strength for recognition *d'*	$F(1, 52) = 26.18$	< .001*	.33
Study strength for classification *d'*	$F(1, 52) = 5.68$	.02*	.10
Study strength by effect type	$F(1, 52) = 11.91$	< .001*	.19
Study strength for episodic *d'*	$F(1, 52) = 25.21$	< .001*	.33
Study strength for structural *d'*	$F(1, 52) = 1.29$	.26	
Deadline by effect type	$F(1, 52) = 6.54$	.01*	.11
Deadline for episodic *d'*	$F(1, 52) = 17.56$	.001*	.25
Deadline for structural *d'*	$F < 1$	-	
Effect type by study strength by task type	$F(1, 52) = 8.19$	.01*	.14
Study strength for recognition episodic effect	$F(1, 52) = 30.04$	< .001*	.37
Study strength for classification episodic effect	$F < 1$	-	-
Study strength for structural effect in both task types	Highest $F(1, 52) = 2.80$	.10	-
There were no other significant effects	Highest $F(1, 52) = 2.31$	.13	-

Note - Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

The main effect of deadline indicated that there was less overall discrimination in the short-deadline condition ($M = 0.48$, $SE = .04$) then in the long-deadline condition ($M = 0.69$, $SE = .04$). The effect-type main-effect indicated that the episodic *d'* ($M = 0.95$, $SE = .07$) was larger than the structural *d'* ($M = 0.22$, $SE = .04$). The study-strength main effect indicated better overall discrimination in the strong-study condition ($M = 0.78$, $SE = .05$) than in the weak-study condition ($M = 0.39$, $SE = .05$).

The study-strength by task-type interaction indicated that recognition task overall *d'* increased from weak-study ($M = 0.35$, $SE = .07$) to strong-study ($M = 0.89$, $SE = .08$) conditions, and classification task overall *d'* also increased from weak-study ($M = 0.43$, $SE = .07$) to strong-study ($M = 0.68$, $SE = .07$) conditions. However, the increase was greater in the recognition task than in the classification task.

The study-strength by effect-type interaction indicated that the episodic *d'* increased from weak-study ($M = 0.61$, $SE = .09$) to strong-study ($M = 1.30$, $SE = .10$) conditions, whilst the structural *d'* did not increase from weak-study ($M = .17$, $SE = .06$) to strong-study ($M = 0.27$, $SE = .06$) conditions.

The deadline by effect-type interaction indicated the episodic *d'* was smaller in the short-deadline ($M = 0.74$, $SE = .08$) than in the long-deadline condition ($M = 1.16$, $SE = .09$) whilst there was no change in the structural *d'* from the short-deadline ($M = 0.22$, $SE = .05$) to the long-deadline condition ($M = 0.22$, $SE = .06$).

The means for the three-way effect-type by study-strength by task-type interaction can be seen in the two right-hand columns of Table 4.1. The episodic effect increased from weak-study to strong-study conditions in recognition, but did not change in classification. The structural effect did not change by study strength in either recognition or classification.

4.2.3.2 Analysis of endorsement rates.

Endorsement rates were each entered into a 2 x 2 x 2 x 2 ANOVA with between-subject factors of study strength (weak study versus strong study) and task type (recognition versus classification) and within-subject factors of deadline (short versus long) and word type (old versus NC versus NI). See Table 4.3 for means and standard errors and Table 4.4 for the results of the ANOVA and related pairwise comparisons.

Table 4.3

Endorsement Rates by Study Strength, Word Type, Task Type, and Deadline from Experiment 7 (SE in brackets)

Task and word type	Short deadline		Long deadline		Total	
	Weak study	Strong study	Weak study	Strong study	Weak study	Strong study
Recognition						
Old	.63 (.05)	.74 (.04)	.59 (.04)	.75 (.03)	.61 (.04)	.75 (.04)
NC	.51 (.05)	.31 (.05)	.37 (.04)	.16 (.03)	.44 (.05)	.23 (.05)
NI	.45 (.05)	.22 (.04)	.29 (.04)	.12 (.02)	.37 (.05)	.17 (.05)
Classification						
Old	.62 (.06)	.66 (.06)	.58 (.05)	.75 (.04)	.60 (.04)	.70 (.04)
NC	.42 (.05)	.41 (.07)	.26 (.03)	.34 (.07)	.34 (.05)	.37 (.05)
NI	.42 (.06)	.31 (.06)	.23 (.04)	.27 (.07)	.32 (.05)	.29 (.05)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 4.4

Results of 3(Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Effect Type) x 2 (Deadline) ANOVA on Endorsement Rates for Experiment 7

Effect	*F* value (df)	*p* value	η^2
Deadline	$F(1, 52) = 21.23$	< .001*	.29
Word type	$F(2, 104) = 216.99$	< .001*	.81
Old versus NC word type	$F(1, 55) = 130.25$	< .001*	.70
NC versus NI word type	$F(1, 55) = 25.14$	< .001*	.31
Deadline by study strength	$F(1, 52) = 4.34$	.04*	.08
Deadline in weak condition	$F(1, 52) = 24.17$	< .001*	.32
Deadline in strong condition	$F(1, 52) = 2.96$	.09	-
Word type by study strength	$F(2, 104) = 22.80$	< .001*	.30
Study strength for old words	$F(1, 52) = 8.95$	.004*	.15
Study strength for NC words	$F(1, 52) = 3.36$	.07	-
Study strength for NI words	$F(1, 52) = 5.56$	.02*	.10
Deadline by word type	$F(2, 104) = 16.23$	< .001*	.24
Deadline for old words	$F < 1$	-	-
Deadline for NC words	$F(1, 52) = 28.49$	< .001*	.35
Deadline for NI words	$F(1, 52) = 33.17$	< .001*	.39
Word type by study-strength by task type	$F(2, 104) = 6.58$	.002*	.11
Study strength for recognition old words	$F(1, 52) = 5.46$	.02*	.09
Study strength for recognition NC words	$F(1, 52) = 9.07$	.004*	.15
Study strength for recognition NI words	$F(1, 52) = 7.84$	.007*	.13
Study strength for classification old words	$F(1, 52) = 3.55$	.06	-
Study strength for classification NC words	$F(1, 52) = .24$	.63	-
Study strength for classification NI words	$F(1, 52) = .24$	.63	-
There were no other significant effects	$F(1, 52) = 2.40$	.13	-

Note - Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

The main effect of deadline reflected more endorsements in the short deadline condition ($M = .47$, $SE = .02$) than in the long deadline condition ($M = .39$, $SE = .02$). The main effect of word type reflected more old endorsements ($M = .66$, $SE = .02$) than NC

endorsements (*M* = .35, *SE* = .02) and more NC endorsements than NI endorsements overall (*M* = .29, *SE* = .02).

The deadline by study-strength interaction reflected the fact that in the weak-study condition there were fewer endorsements with a short deadline (*M* = .51, *SE* = .03) than with a long deadline (*M* = .39, *SE* = .03) while in the strong-study condition there was no difference between short deadline (*M* = .44, *SE* = .04) and long deadline (*M* = .40, *SE* = .03).

The word-type by study-strength interaction was due to there being more old endorsements in the strong-study condition (*M* = .73, *SE* = .03) than in the weak-study condition (*M* = .60, *SE* = .03) and fewer NI endorsements in the strong-study (*M* = .23, *SE* = .03) than in the weak-study condition (*M* = .35, *SE* = .03). There was no difference in the NC endorsements between weak-study (*M* = .39, *SE* = .03) and strong-study (*M* = .30, *SE* = .03) conditions.

The deadline by word-type interaction reflected no difference in old endorsements between short deadline (*M* = .66, *SE* = .03) and long deadline (*M* = .67, *SE* = .02) whilst there were more NC endorsements with a short deadline (*M* = .41, *SE* = .03) than with a long deadline (*M* = .23, *SE* = .02) and more NI endorsements with a short deadline (*M* = .35, *SE* = .03) than with a long deadline (*M* = .23, *SE* = .02).

The means and standard errors for the word-type by study-strength by task-type interaction are shown in Table 4.3. The interaction was due to there being no difference in the endorsement rates due to study strength in classification, whilst in recognition old endorsements increased from weak- to strong-study conditions whilst both NC and NI endorsements decreased from weak- to strong-study conditions.

4.2.4 Discussion.

The central result here is that even under short-deadline conditions the episodic effect still increased with study strength. It could be that the episodic effect increase in the recognition condition was due to residual recollection. This is unlikely as Jacoby (1999) found that the same deadline manipulation reduced recollection and allowed an increase in familiarity to be measured, so at least some of the increase in the episodic effect in the recognition condition is likely to be due to episodic familiarity. In Experiment 7, the overall episodic effect was smaller in the short-deadline condition than in the long deadline. Thus, consistent with Jacoby's results, it is likely that in the short-deadline condition episodic familiarity would have played a greater role than in the long-deadline condition. As the

episodic effect still increased with study strength under a short deadline, a large proportion of this increase is likely to be due to episodic familiarity.

Although some studies may suggest that familiarity does not increase with repetitions, this is misleading and is often the result of directly equating know responses to familiarity. For instance, Gardiner, Kaminska, Dixon and Java (1996) found that know responses did not increase with repetitions when using classical music as stimuli. Jacoby, Jones and Dolan (1998) conceptually replicated the experiment and demonstrated that equating know responses with familiarity was misleading, because doing so assumes that familiarity only occurs in the absence of recollection. Assuming that recollection and familiarity can co-occur, Jacoby et al. found that although know responses did not increase with repetition at study, familiarity actually did increase with repetition. This further supports the interpretation that the increase in the episodic effect by study strength was due to an increase in episodic familiarity.

The results from the classification task seem more problematic. Although numerically the episodic *d'* increased, it was not a statistically significant increase. It could be that the episodic effect increase in the recognition condition was due to residual recollection and that this recollection was further diminished in the classification condition. However, another possible explanation for the insensitivity of the episodic effect to study strength in classification is that the task complexity in classification reduced all forms of evidence that a participant might use to perform the task. Anecdotally, participants found the classification version of the experiment much harder than the recognition condition. This is not too surprising – a memory task is easier for participants to perform than a classification task when the rule set is deliberately designed to be hard to discover. Both Whittlesea and Dorken (1993) and Poznanski and Tzelgov (2010) have demonstrated that task demands affect performance. It is likely that switching between deadlines also increased the task difficulty relative to previous experiments. Thus the demanding nature of the task may have acted as an inhibitor, reducing the extent to which the episodic effect was sensitive to study strength.

The structural effect was somewhat more erratic in Experiment 7 than in Experiment 5. Although the structural effect was not increased by study strength, in some conditions the structural effect did not reach above-chance levels. The study phase of Experiment 7 was identical to Experiment 5, which showed a reliable structural effect. The deadline thus seemed to interfere with the expression of the structural effect. In the short-deadline condition of both tasks and the long-deadline condition of the classification task, the

structural effect was above chance in the strong-study condition but not in the weak-study condition. It is possible that a minimum amount of attention or processing time is required for the structural effect to be expressed. The effect of study strength in this case could be to produce a more stable effect that can be more quickly expressed, but not to increase the actual magnitude of the effect. However, in the long-deadline condition of the recognition task the above-chance structural effect appeared in the weak-study condition and not in the strong-study condition. This suggests another two alternative explanations. One is that there was insufficient power to detect the structural effect in this experiment. Given the magnitude of the structural effect this is entirely possible, and thus the fluctuations in the structural effect seen in Experiment 7 could just be error. Another possibility is that instead of making the structural effect less noisy, the deadline manipulation in fact increased the noise in the structural effect making it harder, not easier, to detect. This could have occurred because of the overall task difficulty due to the frequent changes of deadline. Given the stability of the structural effect in Experiment 5, it seems likely that the task difficulty, rather than power problems, were the culprit (Power is discussed further in Chapter 5). Regardless, the fact that the episodic effect increased by study strength whilst the structural effect did not suggests that episodic familiarity does not drive the structural effect. Experiment 8 utilised a different method to reduce recollection in order to try and confirm this result when a more stable structural effect is present.

The endorsement rate data indicated that shifting from a long to a short deadline resulted in an increase in NC and NI endorsements and no change in old endorsements. Although no change in old endorsements may suggest that the deadline did not shift the old distribution, this is not the case. The increase in NC and NI endorsements indicated that participants adopted a more liberal criterion (i.e. the criterion shifted to the left) with a short deadline than with a long deadline. In order for old endorsements to stay static in this case, the old distribution would have to shift to the left (see *Figure 4.2*). Thus the deadline manipulation did indeed reduce the strength of evidence for old words. Although the shift in the old distribution could be due to a reduction in either episodic familiarity or recollection, it would be consistent with the existing literature (Jacoby, 1999) to interpret the shift as evidence that recollection has been impaired.

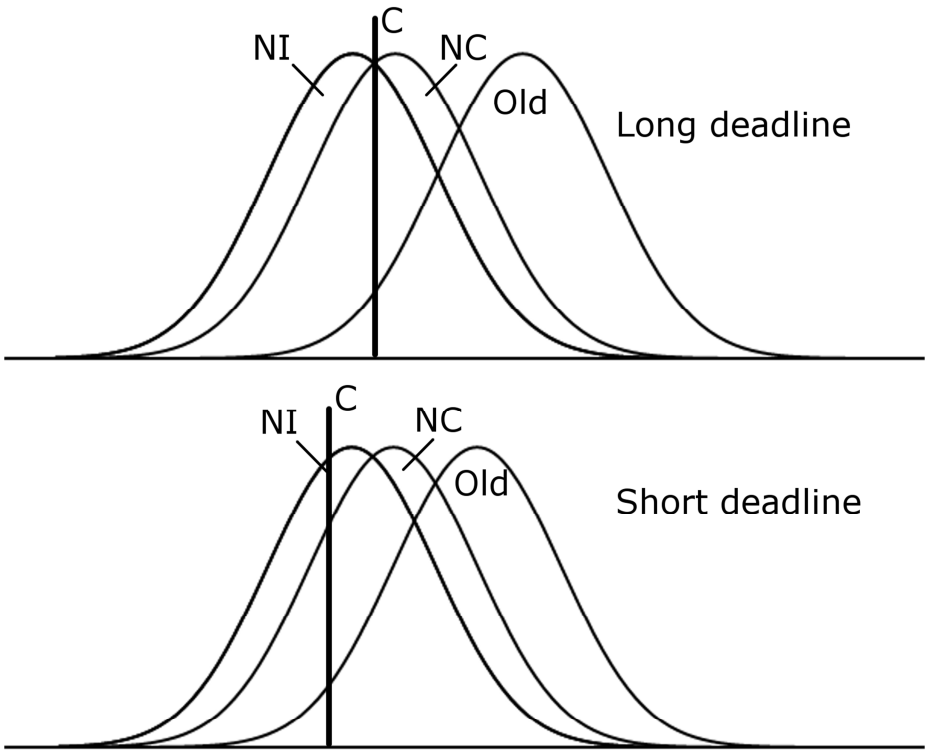

Figure 4.2. Patterns of changes by deadline in Experiment 7. Old = old words; NC = new consistent words; NI = new inconsistent words; C = criterion.

It is worth noting that because deadline was random on each individual trial, deadline-based criterion shifts indicated that participants shifted their criterion on a trial-by-trial basis. Some studies have found that participants do not shift their criterion in this way (e.g. Morrell et al., 2002; Stretch & Wixted, 1998). Rhodes and Jacoby (2007) demonstrated that participants could shift their criterion on a trial-by-trial basis but often saw no reason to do so, requiring feedback to make it clear that there was a reason to shift criterion. Bruno, Higham and Perfect (2009) demonstrated that participants only shifted their criterion in response to a within-participant manipulation of study strength when high task difficulty (such as having short study presentation times) resulted in participants judging the stimuli to have a low global subjective memorability. The results from the current experiment appear to support such conclusions – the deadline manipulation resulted in a difficult task and so participants shifted their criterion in response to the feedback that it would be either a short- or long-deadline trial. Although the participants may not have judged the study list to result in low global subjective memorability it may be that participants' beliefs about how likely they were to remember anything with such short deadlines created a similar effect.

Despite the deadline manipulation a full mirror effect was found in the recognition task. A change from weak- to strong-study conditions resulted in an increase in old endorsements rates and a decrease in NC and NI endorsement rates. There were also more old than NC endorsements and more NC than NI endorsements. Just as in Experiment 5, participants adopted a more conservative criterion in strong-study than in weak-study conditions and the old distribution shifted to the right from weak- to strong-study conditions, at least in the recognition task. Despite recollection being reduced by the deadline manipulation, study strength still shifted the old distribution to the right. Assuming that recollection was reduced to a reasonable degree, this would suggest that study strength was boosting the strength of evidence for old words through episodic familiarity. Additionally, the magnitude of the difference between the NC and NI endorsement rates was unchanged by study strength. Consistent with the results from the *d'* analysis, this suggests that the familiarity supporting the structural effect is indeed not the same thing as the familiarity supporting the episodic effect.

In classification, no mirror effect was found. In fact, study strength did not affect any of the endorsement rates. This is difficult to interpret, as there should be at least some change in NC and NI endorsement rates due to different criteria in weak- and strong-study conditions. It is possible that the classification condition was noisier than the recognition condition due to the increased difficulty of the task, since participants were switching between long and short deadlines frequently. In this case, more power would be needed to detect any changes in the endorsement rates. Power is addressed in Chapter 5.

In conclusion, the episodic effect increased by study strength in both deadline conditions. Participants also seemed to shift their criterion on an item-by-item basis. The structural effect data were less clear. Imposing a deadline seemed to make the structural effect harder to express.

Experiment 8 attempted to address the problems with the classification condition and the unstable structural effect found in Experiment 7. Recollection was reduced using a different method, applied as a between-subject manipulation in order to reduce overall task difficulty. Using a different method to reduce recollection should provide converging evidence of the effect of study strength on episodic and structural familiarity.

4.3 Experiment 8

4.3.1 Introduction.

Experiment 7 had two main aims – to see if the structural effect and the episodic effect reacted differently to study strength when recollection was reduced and to increase the proportion of familiarity that was contributing to both effects. Experiment 8 had the exact same aims as Experiment 7, except a different method was used to reduce recollection. Distraction at test is thought to reduce recollection whilst leaving familiarity intact (Gruppuso, Lindsay, & Kelley, 1997; Jacoby, 1991; Yonelinas, 2002). Thus the odd-number counting task from Experiment 2, Chapter 2 was used at test in order to reduce the effects of recollection, as used by Craik (1982) and Jacoby (1991).

4.3.2 Predictions.

The predictions for Experiment 8 are similar to those in Experiment 7. Distraction should result in participants judging the task as difficult and adopting a more liberal criterion to compensate. This will result in more NI and NC endorsements when distracted than when not distracted. Additionally, distraction should reduce recollection, shifting the old distribution to the left from not-distracted to distracted conditions resulting in either a decrease or no change in old endorsements (depending on how much the criterion shifts).

Also of interest is the effect of study strength in the distracted condition. Distraction reduces recollection but does not affect episodic familiarity (Gruppuso et al., 1997; Yonelinas, 2002). Of central interest is whether the episodic and structural effects react the same way to study strength in the distracted condition. The increase in study strength should shift the criterion to the right as before. An increase in study strength when distracted will result in the old distribution shifting to the right if study strength increases episodic familiarity, resulting in an increase, or no change, in old endorsements depending on the extent of the criterion shift. This will lead to an increase in the episodic effect. If study strength does not increase episodic familiarity, the old distribution will stay static and old endorsements will decrease due to the criterion shift and there will be no increase in the episodic effect. The increase in study strength could shift the NC distribution to the right, resulting in an increase or no change in NC endorsements depending on the criterion shift. The structural effect would increase with study strength in this case. Alternatively, the NC distribution may not be affected by study strength, resulting in a decrease in NC endorsements and no change in the structural effect from weak- to strong-study conditions.

The criterion shift will also result in lower NI endorsement rates in the strong-study condition than in the weak-study condition. If the overall pattern of changes results in the episodic and structural effects reacting the same way to study strength, then it is likely episodic familiarity underlies both effects. If on the other hand the structural and episodic effects react differently to study strength, then it is likely that episodic familiarity does not underlie the structural effect.

4.3.3 Method.

4.3.3.1 Participants.

One hundred and thirty-four undergraduates from the UoS participated in the experiment. All participants were either given course credits or £5 payment.

4.3.3.2 Materials.

The same word lists were used as in Experiment 7, except they were not divided into short and long deadline types as there was no deadline manipulation in Experiment 8.

4.3.3.3 Design.

The design was the same as Experiment 7 except that there was no response deadline imposed. Instead, participants responded to all words at their own pace. Half of the participants were distracted at test whilst the other half were not distracted at test. The distraction task was the same distraction task as used in Experiment 2. Participants heard a stream of numbers and had to press space when they heard three odd numbers in a row. A box appeared to remind participants of this task the first time they failed to identify three odd numbers in a row and then again every four failures after the first. Thus the design used between-subject manipulations of distraction at test (distracted versus not distracted), task type (classification versus recognition) and study strength (weak study versus strong study).

4.3.3.4 Procedure.

The study phase and retention interval were the same as Experiment 7. The test phase was the same except for the following changes. In the test phase there was no deadline – the experiment was self-paced. Participants made all their responses by clicking radio buttons rather than pressing a button on the keyboard. At test, half of the participants completed their task whilst distracted by the number-counting task whilst the other half of the participants were not distracted at test. Half of the participants were given a recognition task and half

were given a classification task, with the distracted and not-distracted participants being assigned equally to each task. After the test phase, the questionnaire was again administered as in Experiment 7.

4.3.4 Results.

Seven participants were excluded from the analysis: Six because they did not perform the distraction task properly and one because they correctly selected the individual elements of the conjunctive rule-set on the questionnaire, although they did not correctly identify the nature of the conjunction.

4.3.4.1 Analysis of episodic and structural d'.

As in previous experiments, the *d'* measures were compared with chance performance using 95% confidence intervals. The results can be seen in Table 4.5 below. The changes in magnitude of the *d'* measures were investigated with a 2 x 2 x 2 x 2 ANOVA with between-participant factors of task type (recognition versus classification), study strength (weak-study versus strong-study) and distraction (distracted versus not distracted) and a within-subject factor of effect type (episodic versus structural). Results and pairwise comparisons can be seen in Table 4.6.

Table 4.5

d' by Study Strength, Task Type, Effect Type and Distraction from Experiment 8 (SE in Brackets)

Task and *d'* type	Not Distracted		Distracted		Total	
	Weak study	Strong study	Weak study	Strong study	Weak study	Strong study
Recognition						
Episodic *d'*	.94 (.11)*	2.40 (.20)*	.84 (.09)*	2.08 (.23)*	.89 (.10)*	2.24 (.11)*
Structural *d'*	.28 (.06)*	.28 (.10)*	.32 (.10)*	.23 (.10)*	.30 (.06)*	.26 (.06)*
Classification						
Episodic *d'*	.76 (.09)*	1.29 (.15)*	.89 (.08)*	1.55 (.20)*	.83 (.11)*	1.42 (.11)*
Structural *d'*	.14 (.07)*	.42 (.09)*	.30 (.07)*	.32 (.10)*	.22 (.06)*	.37 (.06)*

* = Lower bound of 95% confidence interval above chance level of 0.

Table 4.6

Results of 2 (Study Strength) x 2 (Task Type) x 2 (Effect Type) x 2 (Distraction) ANOVA on d' for Experiment 8

Effect	*F* value (df)	*p* value	η^2
Task type	$F(1, 119) = 13.68$	< .001*	.10
Effect type	$F(1, 119) = 251.60$	< .001*	.68
Study strength	$F(1, 119) = 77.37$	< .001*	.39
Effect type by task type	$F(1, 119) = 11.87$	.001*	.09
Task type for episodic *d'*	$F(1, 119) = 16.91$	.001*	.12
Task type for structural *d'*	$F < 1$	-	-
Task type by study strength	$F(1, 119) = 5.86$	.02*	.05
Task type for weak study	$F<1$	-	-
Task type for strong study	$F(1, 119) = 18.64$	< .001*	.13
Effect type by study strength	$F(1, 119) = 47.34$	< .001*	.28
Study strength for episodic *d'*	$F(1, 119) = 29.62$	< .001*	.40
Study strength for structural *d*'	$F < 1$		
Effect type by study strength by task type	$F(1, 119) = 12.67$	< .001*	.10
Study strength for recognition episodic *d'*	$F(1, 119) = 80.97$	.001 *	40
Study strength for recognition structural *d'*	$F < 1$	-	-
Study strength for classification episodic *d'*	$F(1, 119) = 14.38$	.001*	.11
Study strength for classification structural *d'*	$F < 1$	-	-
There were no other significant effects	Highest $F(1, 119) = 3.59$	.09	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

The task-type main effect reflected higher *d'* in recognition ($M = .92$, $SE = .04$) than in classification ($M = .71$, $SE = .04$). The effect-type main effect reflected higher episodic *d'* ($M = 1.34$, $SE = .05$) than structural *d'* ($M = .29$, $SE = .03$). The study-strength main effect reflected higher *d'* in the strong-study condition ($M = 1.07$, $SE = .04$) than in the weak-study condition ($M = .56$, $SE = .04$).

The effect-type by task-type interaction reflected the fact that the structural *d'* was the same in recognition ($M = .28$, $SE = .04$) and in classification ($M = .30$, $SE = .04$), whilst the episodic *d'* was greater in recognition ($M = 1.57$, $SE = .07$) than in classification ($M = 1.12$, $SE = .08$).

The study-strength by task-type interaction reflected the fact that there was no difference between the classification *d'* ($M = .52$ $SE = .06$) and the recognition *d'* ($M = .60$,

SE = .06) in the weak-study condition whilst in the strong-study condition the recognition *d'* (*M* = 1.24, *SE* = .06) was greater than the classification *d'* (*M* = .89, *SE* = .06).

The effect-type by study-strength interaction was due to the episodic *d'* increasing from weak-study (*M* = .87, *SE* = .08) to strong-study conditions (*M* = 1.83, *SE* = .08), whilst the structural *d'* did not change from weak-study (*M* = .26, *SE* = .04) to strong-study conditions (*M* = .31, *SE* = .04).

The means and standard errors for the effect-type by study-strength by task-type interaction can be seen in Table 4.5. This interaction reflected the fact that the magnitude of the increase of the episodic *d'* from weak- to strong-study conditions was greater in recognition than in classification, whilst the structural *d'* did not change by study strength or task type.

4.3.4.2 Analysis of endorsement rates.

Endorsement rates were entered into a 3 x 2 x 2 x 2 ANOVA with a within-subject factor of word type (old versus NC versus NI) and between-subject factors of study strength (weak-study versus strong-study), task type (recognition versus classification) and distraction condition (distracted versus not-distracted). For means and standard errors see Table 4.7 and for the results of the ANOVA and related pairwise comparisons see Table 4.8.

Table 4.7

Endorsements Rates by Word Type, Task Type and Distraction from Experiment 8 (SE in brackets)

Task and word type	Not Distracted		Distracted		Total	
	Weak study	Strong study	Weak study	Strong study	Weak study	Strong study
Recognition						
Old	.73 (.02)	.87 (.03)	.72 (.02)	.84 (.03)	.72 (.02)	.85 (.02)
NC	.39 (.03)	.17 (.04)	.42 (.03)	.20 (.03)	.40 (.03)	.18 (.03)
NI	.30 (.04)	.12 (.03)	.32 (.04)	.16 (.03)	.31 (.03)	.14 (.03)
Classification						
Old	.76 (.03)	.79 (.03)	.75 (.03)	.81 (.03)	.75 (.02)	.80 (.02)
NC	.50 (.04)	.37 (.03)	.44 (.04)	.31 (.05)	.47 (.03)	.34 (.03)
NI	.44 (.04)	.26 (.04)	.34 (.04)	.23 (.05)	.39 (.03)	.24 (.03)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table 4.8

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Distraction) ANOVA on Endorsement Rates for Experiment 8

Effect	*F* value (df)	*p* value	η^2
Word type	$F(2, 238) = 906.10$	< .001*	.88
Old versus NC	$F(1, 126) = 494.87$	< .001*	.80
NC versus NI	$F(1, 126) = 90.19$	< .001*	.42
Task type	$F(1, 119) = 9.98$	.002*	.08
Study strength	$F(1, 119) = 16.17$	< .001*	.12
Word type by task type	$F(2, 238) = 13.19$	< .001*	.10
Task type for old endorsements	$F < 1$		
Task type for NC endorsements	$F(1, 119) = 17.89$	< .001*	.13
Task type for NI endorsements	$F(1, 119) = 10.69$	.001*	.08
Word type by study strength	$F(2, 238) = 62.50$	< .001*	.34
Study strength for old endorsements	$F(1, 119) = 17.44$	< .001*	.13
Study strength for NC endorsements	$F(1, 119) = 41.95$	< .001*	.26
Study strength for NI endorsements	$F(1, 119) = 31.58$	< .001*	.21
Word type by task type by study strength	$F(2, 238) = 5.92$	.003*	.05
Study strength for recognition old endorsements	$F(1, 119) = 20.11$	< .001*	.14
Study strength for recognition NC endorsements	$F(1, 119) = 35.46$	< .001*	.23
Study strength for recognition NI endorsements	$F(1, 119) = 19.02$	< .001*	.14
Study strength for classification old endorsements	$F(1, 119) = 2.20$	.14	-
Study strength for classification NC endorsements	$F(1, 119) = 10.63$	< .001*	.08
Study strength for classification NI endorsements	$F(1, 119) = 12.98$	< .001*	.10
Word type by task type by distraction	$F(2, 238) = 3.54$	.03*	.03
Task type for not-distracted old endorsements	$F < 1$	-	-
Task type for not-distracted NC endorsements	$F(1, 119) = 17.03$	< .001*	.12
Task type for not-distracted NI endorsements	$F(1, 119) = 12.84$	< .001*	.10
Task type for distracted old endorsements	$F < 1$	-	-
Task type for distracted NC endorsements	$F(1, 119) = 3.55$	.06	-
Task type for distracted NI endorsements	$F(1, 119) = 1.15$	.28	-
No other effects were significant	$F(1, 119) = 1.51$	.22	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

The word-type main effect reflected more old endorsements (*M* = .78, *SE* = .01) than NC endorsements (*M* = .35, *SE* = .01) and more NC endorsements than NI endorsements (*M* = .27, *SE* = .01). The task-type main effect reflected more endorsements in classification (*M* = .50, *SE* = .01) than in recognition (*M* = .44, *SE* = .01). The study-strength main effect reflected more endorsements in weak study (*M* = .51, *SE* = .01) than in strong study (*M* = .42, *SE* = .01).

The word-type by task-type interaction reflected no difference in old endorsements between recognition (*M* = .79, SE = .01) and classification (*M* = .78, SE = .01), while there were more NC endorsements in classification (*M* = .41, *SE* = .02) than in recognition (*M* = .29, *SE* = .02) and more NI endorsements in classification (*M* = .32, *SE* = .02) than recognition (*M* = .22, *SE* = .02).

The word-type by study-strength interaction reflected the fact that old endorsements increased from weak-study (*M* = 74, *SE* = .01) to strong-study (*M* = .82, *SE* = .01) conditions whilst NC endorsements decreased from weak-study (*M* = .44, *SE* = .02) to strong-study (*M* = .35, *SE* = .02) conditions and NI endorsements also decreased from weak-study (*M* = .35, *SE* = .02) to strong-study (*M* = .19, *SE* = .02) conditions.

The means and standard errors for the word-type by task-type by study-strength interaction can be seen in Table 4.7. This interaction reflected the fact that old endorsements increased from weak- to strong-study conditions in recognition but not in classification. Both NC and NI endorsements decreased from weak- to strong-study conditions in recognition and classification.

The word-type by task-type by distraction interaction reflected the fact that in not-distracted conditions NC endorsements increased from recognition (*M* = .28, *SE* = .03) to classification (*M* = .43, *SE* = .03) and NI endorsements increased from recognition (*M* = .21, *SE* = .03) to classification (*M* = .35, *SE* = .03) whilst old endorsements did not change from recognition (*M* = .80, *SE* = .02) to classification tasks (*M* = .77, *SE* = .02). In the distracted condition there was no difference in old endorsements from recognition (*M* = .78, *SE* = .02) to classification (*M* = .78, *SE* = .02), no difference in NC endorsements from recognition (*M* = .31, *SE* = .03) to classification (*M* = .38, *SE* = .03) and no difference in NI endorsements from recognition (*M* = .24, *SE* = .03) to classification (*M* = .28, *SE* = .03).

4.3.5 Discussion.

The results of the *d'* analysis were disappointing in that there were no effects of distraction. The pattern of data was otherwise similar to Experiment 5. The episodic effect

was sensitive to study strength whilst the structural effect was not sensitive to study strength. The episodic effect was smaller in classification and increased less with study strength in classification. The reduced episodic effect in classification was consistent with participants using less recollection in classification than in recognition, just as in Experiment 5. The structural effect was above chance in every condition in the current experiment, which further supports the interpretation that the low structural effects in Experiment 7 were due to the deadline manipulation introducing more noise. However, see Chapter 5 for more discussion on power.

As in Experiment 5, an increase in study strength resulted in increased old endorsements and decreased NC and NI endorsements. There was a single effect of distraction in the endorsement data. When participants were not distracted, NC and NI endorsements increased from recognition to classification whilst old endorsements did not change. This suggests that, as in earlier experiments, participants adopted a more liberal criterion in classification and the old distribution was lower on the strength-of-evidence scale in classification than in recognition. However, when distracted there was no difference in any endorsement rates from recognition to classification. Distraction appears to have stopped participants adopting a more liberal criterion in classification and also appears to have prevented the drop in the strength of evidence of the old distribution from recognition to classification. Why this could be is not apparent from the data. It is possible that the distraction manipulation resulted in participants spending longer on each trial, allowing the recollection in the classification task to reach the same levels as in the recognition task. This additional recollection could also explain the lack of a criterion shift if participants used an estimate of their recollection as a basis for setting their criterion. Overall, it appears that the distraction manipulation failed to reduce recollection in this experiment.

There are several candidate explanations for why the distraction manipulation did not reduce recollection. Lozito and Mulligan (2010) used a variety of secondary tasks on a recall test which varied in the match to the primary task of the materials used and to the primary task and also in the response frequency required. In their Experiment 4, the three-odd-numbers distraction task did not reduce recall performance because the primary task was word-based and the response frequency of the three-odd-numbers task was low. Although these data were gathered using a recall task, both of these explanations could apply to Experiment 8. A response was only required when three odd numbers occurred, so the response frequency was low. The secondary task used numbers whereas the primary task used words, so the materials were mismatched. A second explanation comes from Hicks and

Marsh (2000), who used several different secondary tasks in a recognition test. They concluded that some secondary tasks require insufficient sustained attention to reduce recollection. For instance, when using a modified form of the digit-load task from Baddeley, Lewis, Eldrige and Thomson (1984) in which participants had to listen to and recite a sequence of numbers, there was no detriment to recognition performance. However, when a metronome was used to force participants to respond at specific time intervals, the digit-load task did reduce recognition performance. Hicks and Marsh concluded that adhering to external pacing was the part of the task that demanded the most attention. Thus it is also possible that the three-odd-number task in Experiment 8 simply did not demand enough attention to reduce recognition performance. One final possibility is that the self-paced nature of Experiment 8 allowed participants to stop the primary task when they heard two odd numbers, and then resume it upon hearing the next number. Although time taken to complete the experiment was not recorded, the distracted condition did seem to take longer than the not-distracted condition. Though no further distraction experiments are presented here, future research should use a distraction task that requires frequent responses, dictated by a metronome, involving words.

4.4 Conclusions

Experiments 1 to 8 have demonstrated that the structural effect appears to be resistant to study-strength manipulations, instructions manipulations and distraction. Although there seem to be ways to suppress the expression of the structural effect using deadlines, it appears to be difficult to increase the magnitude of an expressed effect. It could be that the structural effect is not sensitive to study strength because it is a small and noisy effect, increases in which are hard to detect. Chapter 5 addresses this criticism with some additional analysis of the data.

If the structural effect is driven by some kind of structural familiarity, then it would appear that this is a different beast to episodic familiarity, assuming that episodic familiarity increases with study strength as is specified by the existing literature. If so, then the question remains, if structural familiarity underlies the structural effect, what underlies structural familiarity? Direct abstraction of the underlying rule set could be responsible for structural familiarity. However, other factors such as chunk strength could also contribute to structural familiarity. Alternatively, structural familiarity that is insensitive to study strength might emerge from the way that existing models of memory operate, perhaps through a chorus-of-instance type mechanism such as that used by MINERVA. Thus Chapter 5 looks at what

memory models have to offer in terms of the insensitivity of the structural effect to study strength.

This page intentionally left blank

5 Chapter 5: Limitations and Simulations

5.1 Overview

This chapter focuses on two areas. Firstly, it addresses some limitations of the previous studies. Specifically, that the structural effect could just be a small and noisy effect; that the studies lacked statistical power; and that the structural effect could be due to the surface characteristics of the stimuli expressed through chunk-strength. Secondly it explores recognition memory models. Some explanations for rule learning in implicit learning revolve around chorus-of-instance type effects (e.g. Vokey & Brooks, 1992). Computational models of recognition memory such as MINERVA 2 (Hintzman, 1984, 1986) use similar mechanisms. That being the case, they may also be able to explain the structural effect. Thus recognition memory models are reviewed and MINERVA simulations conducted to test if the structural effect occurs as a natural extension of how the models operate.

5.2 Limitations

5.2.1 Small and noisy effect.

One limitation of the experiments is that it is possible that the structural effect is of a small magnitude with a great deal of variation. Small changes in the structural effect in response to study strength might be masked by the general background variation. Attempts were made in Chapter 3 to test this by boosting the magnitude of the structural effect. As these attempts were unsuccessful, another approach was taken to see if a larger structural effect is in fact sensitive to study strength. The data from Experiments 4 and 5 were split into equal thirds by the magnitude of the structural effect, such that the top third contained the largest structural effects, then the middle third contained the next largest structural effects and the final third contained the lowest structural effects. If the structural effect is in fact sensitive to study strength, then the top third should be sensitive to an increase in study strength as this section of the data comprises the largest structural effects.

5.2.1.1 Results.

Structural effects in the weak- and strong-study conditions were compared with each other using *t* tests. The structural effect was insensitive to study strength in all cases – see Table 5.1 for means and standard errors and Table 5.2 for the results of the t-tests.

Table 5.1

Magnitude Spilt of Structural Effects by Task Type and Study Strength from Experiments 4 and 5 (SE in brackets)

Task type	Experiment 4		Experiment 5	
	Weak study	Strong study	Weak study	Strong study
Recognition				
Bottom third	-.01 (.08)	.02 (.05)	-.06 (.14)	.02 (.03)
Middle third	.28 (.03)	.32 (.02)	.28 (.06)	.41 (.13)
Top Third	.60 (.06)	.69 (.14)	.74 (.15)	.72 (.15)
Classification				
Bottom Third	-.31 (.07)	-.13 (.11)	-.03 (.08)	-.05 (.03)
Middle Third	.22 (.08)	.14 (.03)	.36 (.02)	.24 (.05)
Top Third	.59 (.06)	.42 (.12)	.67 (.12)	.83 (.14)

Table 5.2

Results of T Test Comparison of Study Strength Conditions for Split Structural Effects

Task type	Experiment 4	Experiment 5
Recognition		
Bottom third	Absolute $t(8) < 1$	Absolute $t(8) < 1$
Middle third	$t(8) = -1.31$, $p = .23$	Absolute $t(8) < 1$
Top Third	Absolute t(8)<1	Absolute $t(8) < 1$
Classification		
Bottom Third	$t(6) = 1.27$, $p = .25$	Absolute $t(8) < 1$
Middle Third	Absolute $t(8) < 1$	$t(10) = 2.18$, $p = .054$
Top Third	$t(6) = -1.403$, $p = .21$	Absolute $t(8) < 1$

5.2.1.2 Discussion.

Even when the participants with the largest structural effects were separated out, the structural effect was still insensitive to study strength. This suggests that it is some property of the structural effect that renders it insensitive to study strength rather than it being a small

and noisy effect. To further test this idea, the next section combines several experiments into one data set.

5.2.2 Power.

In the previous experiments the main results concerned the effects of study strength on two *d'* measures – the episodic *d'* represented performance due to episodic factors and the structural *d'* represented performance due to structural factors. The central analysis was an ANOVA which indicated that the structural *d'* did not increase from weak- to strong-study conditions whilst the episodic *d'* did increase. However, it is possible that the structural *d'* did in fact increase by study strength but only by a very small amount. In this case, the previous experiments may not have had enough participants to provide the power to detect this change. For example, Experiment 3 could detect an effect size of .28 with a power of .95[5]. Assuming a small effect size, it would be preferable if the smallest detectable effect size was less than .20. This is especially relevant as the lack of increase in the structural effect constitutes a null effect. In order to improve the smallest detectable effect size, several experiments were combined into one analysis. Experiments 3, 4 and 5 were chosen for this analysis as they all manipulated study strength and had similar test phases. Experiments 1, 2, and 6 were excluded because they did not manipulate study strength. Experiments 7 and 8 were excluded because they introduced deliberate distractions at test. Combining experiments 3, 4 and 5 yielded 151 participants, resulting in power of .95 to detect an effect size of .15 and power of .80 to detect an effect size of .11.

The data were entered into a single ANOVA with between-participants factors of Experiment (3, 4 or 5), task (classification or recognition), study strength (weak-study or strong-study) and a within-participant factor of effect type (structural or episodic). The primary interest in this analysis was whether the structural effect still did not change with study strength and if this was stable across experiments. Thus only effects involving study strength or experiment are reported.

The analysis yielded a main effect of experiment $F(2, 139) = 44.57$, $p < .001$, $\eta^2 = .39$. There were interactions between effect type and study strength, $F(1, 139) = 18.63$, $p < .001$, $\eta^2 = .12$, effect type and experiment $F(2, 139) = 31.35$, $p < .001$, $\eta^2 = .31$ and study strength and experiment $F(2, 139) = 3.69$, $p = .03$, $\eta^2 = .05$. The study strength and experiment

[5] All power calculations were conducted using sensitivity analysis on G*Power 3.1.2 (Faul, Erdfelder, Lang, & Buchner, 2007)

interaction is ignored as the primary interest is the differential effect of study strength on the two types of effects and this interaction collapses across effect type.

Of most interest is the interaction between effect type and study strength. Pairwise comparisons indicated that this interaction was due to an increase in the episodic effect from weak-study (M = 1.28, SE = .08) to strong-study conditions (M = 1.88, SE = .08), $F(1, 139) = 25.35$, $p < .001$, $\eta^2 = .15$, whilst the structural effect did not change from weak (M = .27, SE = .04) to strong conditions (M = .25, SE = .04), $F < 1$. In other words, the structural effect was not sensitive to study strength even in this high power analysis. This pattern was stable across all three experiments as indicated by the lack of an experiment by study-strength by effect-type interaction.

The interaction between effect type and experiment indicated that the episodic effect varied by experiment, $F(2, 139) = 46.73$, $p < .001$, $\eta^2 = .40$, (Experiment 3 M = 2.42, SE = .11; Experiment 4 M = 1.07, SE = .11; Experiment 5 M = 1.26, SE = .09), whereas the structural did not, $F(2, 139) = 1.51$, $p = .22$ (Experiment 3 M = .26, SE = .06; Experiment 4 M = .19, SE = .06; Experiment 5 M = .32, SE = .05). This is evidence that the structural effect seems insensitive to changes in learning conditions, which was the main difference between these three experiments.

This analysis had high power to detect small effects. The results strongly support the original conclusions of Experiments 3, 4 and 5 – that the structural effect really does not increase from weak- to strong-study conditions. The lack of sensitivity of the structural effect to study strength was not due to a lack of power in the experiments.

5.2.3 Chunk strength.

Many studies have shown that chunk strength can be responsible for participants discriminating between grammatical and non-grammatical AG stimuli (e.g. Dienes et al., 1991; Dulany et al., 1984). Jamieson and Mewhort (2009a) demonstrated that in an AG experiment, rule-consistent strings had certain constraints upon them that created regularities in the frequency with which certain letters or letter strings appeared in specific positions. These contingencies could be used to distinguish between grammatical and non-grammatical strings without any knowledge of the underlying rule set. However, other studies demonstrate that chunk strength alone cannot account for AG performance (Higham, 1997a). The question then is whether chunk strength could account for the structural effect found in Chapters 3 and 4.

Although the stimuli were all natural words it is certainly reasonable to suspect chunk strength could play a role. For instance, more abstract words than concrete words could end in "ly" and more concrete words than abstract words could end "le". If similar restrictions apply to rare and common words then the conjunction of specific chunks could explain why participants endorse more NC words than NI words. Even if participants are capable of learning the underlying rule set, they might not do so if chunk strength is sufficient to perform the task.

In order to test this possibility, a computer programme was written using the Revolution application that mapped all of the chunks in the study stimuli from Experiment 5, along with their frequency of appearance. Two types of chunks were mapped – bigrams (pairs of letters) and trigrams (triplets of letters). For instance, take the word "table". Table has four bigrams in it: "ta", "ab", "bl" and "le"; and three trigrams – "tab", "abl" and "ble". The number of times that each of these bigrams and trigrams occurred across all the items in the study list was computed and logged. The position of the bigrams and trigrams was not taken into account. The test stimuli could then be turned into values representing their chunk strengths. For instance if "ta" had occurred 20 times in the study list and "le" had appeared 15 times in the study list then the word "tale" would have a bigram chunk strength of 35. This was done for both weak- and strong-study conditions and for both study lists.

These strengths were then analysed. In order for chunk strength to be an adequate explanation of the structural effect three criteria must be fulfilled:

- The chunk-strength analysis must predict the existence of a structural effect by NC words having higher chunk strength than NI words.
- The effect must be predicted for both possible rule sets (i.e. for both types of study list).
- The effect must also be insensitive to study strength.

The results reported here represent the total chunk strength of each stimulus. Chunk strength of a test item was taken to be the total number of times that chunks in the test item occurred in the study list. Analyses were also run for maximum chunk strength (only the highest frequency chunk contributes to chunk strength for the stimulus) and average chunk strength (the average of the frequencies of each chunk in the test stimuli is the chunk strength). These results are not presented here as the patterns of results did not differ from those obtained with total chunk strength.

5.2.3.1 Bigram chunk-strength analysis.

See Table 5.3 for the bigram chunk-strengths for NC and NI stimuli in Experiment 5, split down by study strength and rule set i.e. common-concrete/rare-abstract (CCRA) and rare-concrete/common-abstract (RCCA).

Table 5.3

Total Bigram Chunk Strength by Rule Set, Word Type and Study Strength from Experiment 5 (SE in brackets)

Rule set and word type	Weak study	Strong study
CCRA		
NC	19.52 (2.03)	97.62 (10.14)
NI	15.80 (1.66)	79.00 (8.31)
Total	17.66 (4.73)	88.32 (4.73)
RCCA		
NC	14.77 (1.31)	73.87 (6.57)
NI	16.87 (1.61)	84.37 (8.04)
Total	15.74 (3.74)	79.12 (3.74)

Note. NC = New rule-consistent words, NI = New rule-inconsistent words, CCRA = common-concrete/rare-abstract, RCCA = rare-concrete/common-abstract.

The analysis was conducted only on the NC and NI stimuli as the differences in endorsement rates between these stimuli create the structural effect. A separate ANOVA was conducted for each rule set with study strength (weak-study versus strong-study) and word type (NC versus NI).

For CCRA there was a main effect of study strength $F(1, 156) = 111.66$, $p < .001$, $\eta^2 = .42$, and no other effects, highest $F(1, 156) = 2.79$, $p = .10$. The results were the same for RCCA - a main effect of study strength $F(1, 156) = 142.90$, $p < .001$, $\eta^2 = .48$ but no other effects or interactions, highest $F(1, 156) = 1.42$, $p = .24$. The study strength effects reflected the increase in chunk strength from weak- to strong-study conditions – for means and standard errors see Table 5.3. Chunk strength is obviously going to be higher in the strong study condition because the study items are repeated five times in the strong study condition, directly translating to a five-fold increase in chunk strength.

5.2.3.2 Trigram chunk strength analysis.

See Table 5.4 for the trigram chunk-strengths for NC and NI stimuli in Experiment 5.

Table 5.4

Total Trigram Chunk Strength by Rule Set, Word Type and Study Strength from Experiment 5 (SE in brackets)

Rule set and word type	Weak study	Strong study
CCRA		
NC	2.50 (0.46)	12.50 (2.33)
NI	2.07 (0.44)	10.37 (2.18)
Total	2.29 (1.15)	111.44 (1.15)
RCCA		
NC	1.25 (0.30)	6.25 (1.51)
NI	1.77 (0.33)	8.87 (1.67)
Total	1.51 (0.81)	7.56 (0.81)

Note. NC = New rule-consistent words, NI = New rule-inconsistent words, CCRA = common-concrete/rare-abstract, RCCA = rare-concrete/common-abstract.

The data were analysed as for the bigram chunk data. For CCRA there was a main effect of study strength $F(1, 156) = 31.65$, $p < .001$, $\eta^2 = .17$ and no other effects or interactions, all $Fs < 1$. For RCCA there was a main effect of study strength $F(1, 156) = 27.78$, $p < .001$, $\eta^2 = .15$ and no other effects or interactions, highest $F(1, 156) = 1.88$, $p = .17$.

5.2.3.3 Discussion.

For neither bigrams of trigrams was the NC chunk-strength greater than the NI chunk-strength. In fact, numerically the RCCA NI words had a higher chunk-strength than the RCCA NC words. Although this was not statistically significant, it is hard to see how participants might use chunk strength to make their decisions and still produce a structural effect when with a RCCA study-list NI chunk-strength is higher than NC chunk-strength and for a CCRA study-list NI chunk-strength is lower than NC chunk-strength. Even in the classification task where participants were required to distinguish NC words from NI words,

a knowledge of chunk strength would not allow them to perform the task – in fact it may even lead to chance performance overall if data from both rule sets were combined. The classification task creates conditions where participants are required to utilise at least some knowledge of the conjunctive rule-set in order to successfully complete the task. This is consistent with Higham (1997a) in which chunks did not explain performance in several experiments and also with Higham and Brooks (1997) in which participants learnt a conjunctive rule-set.

5.3 Memory Models and Simulations

There are many computational models of memory in the recognition literature such as MINERVA (Hintzman, 1984, 1986), REM (Shiffrin & Steyvers, 1997) and the Strength of Activation Model (SAM - Gillund & Shiffrin, 1984). In this section, several of the popular models are reviewed and their predictions for the structural effect are discussed. Simulations were run using one of the models to investigate whether it could predict the current data.

Two main classes of memory models are be discussed. Global-memory models (GMM) rely on the match of a stimulus to all of the information in memory, much like the chorus-of-instance approach. Likelihood models use a similar mechanism but utilise a likelihood ratio to judge the degree of match. ACT-R is also briefly discussed (Anderson, Bothell, Lebiere, & Matessa, 1998).

5.3.1 Global-memory models.

GMMs all share one common mechanism. Each of them in some way compares a test stimulus to all or some items in memory. Usually this is simplified to mean all the study items in memory. The two GMMs to be discussed here are MINERVA 2 (Hintzman, 1984, 1986) and SAM (Gillund & Shiffrin, 1984).

5.3.1.1 SAM.

SAM assumes that there is only long-term memory, used for information storage, and short-term memory, used for coding and transferring of information to long-term memory. Long-term memory is composed of images, each image being a set of features which contain information. Each image can contain contextual features that are related to the setting in which an image occurred, item information about the specific item such as its name and also inter-item information that links one image to another. Each feature in an image can be activated and will lead to different levels of activation depending on the conditions under

which the feature was stored. When an item is memorised in a recognition task study-phase, SAM models the storage of several different features in an image for that item. For every unit of time that an item is studied the image gains strength in three ways:

- The association of the item being studied to its context results in context strength – that the word is on the study list and is being studied in a laboratory are both examples of context. The context strength depends only on the study time for each word, and the model does not distinguish between different ways of increasing study time. For instance, repetition simply increases the total study time for a particular word. Although in early papers the possibility of new images being created due to repetitions was considered (Gillund & Shiffrin, 1984), later implementations of SAM tend to assume that no new images are created and repetition simply updates existing images.
- The association of the item being studied to other words in short-term memory at the same time. That is, the association of the fourth word on a study list will be strong to the preceding three words on the study list because they are all rehearsed in short-term memory together, whereas it will be weak to the thirteenth word on the study list because by the time it is presented, the fourth word will have been displaced from short-term memory. This depends on how long each word is studied together in memory and on how many words can exist in short-term memory together. If two items were not studied together, they have a low default level of activation – all inter-item features have at least a small amount of strength to account for pre-experimental associations.
- The item being studied and its own image – the extent to which information about the item itself is actually stored.

In a recognition test, the current word and the context in which it is presented are combined and used as a memory probe. The context part of the probe results in a level of activation according to how strongly the context was encoded at study. The word part of the probe results in activation according to how strong the inter-item associations are and on how strongly the item is linked to its own image. The levels of activation also depend on two further factors. As one word can have different meanings depending on its context in a sentence, the probe is subject to a match adjustment depending on how much the test context matches the study context. Also, the activation from each feature is subject to random noise to simulate the fact that memory is fallible. These modified activations are combined and

result in a level of familiarity for that word, which is then compared to a criterion. As words that were studied in the study phase will have high context, high inter-item and high self-strength, the familiarity level associated with old words will often be above this criterion. Factors which may result in recognition failure are context differences, random variance in activations or a strict criterion. One important feature of SAM is that if a word was not seen at study (i.e. a lure in a recognition test) then it will have no context strength and no self-strength because an image for that item does not exist in memory. This results in lures having a level of familiarity defined only by the inter-item association resulting from pre-experimental factors. Because this is subject to random variance, lures sometimes produce a FA. It is important to note that without some level of variability in the activation, lure familiarity would always be lower than target familiarity in a recognition task, and thus the model cannot account for the existence of FARs without this variability. The strength-based mirror effect is explained in this model by a differentiation mechanism (Shiffrin, Ratcliff, & Clark, 1990). Increasing repetitions of study words increases the self-strength and inter-item strength of studied items, increasing the HR. Increasing repetitions also highlights differences between lures and the study items, resulting in a decrease of the FAR.

There are two ways in which SAM could create a structural effect. One way is that the way it encodes concreteness and word frequency could lead naturally to conjunctions. Clues to how SAM might cope with stimuli of different frequencies and concreteness can be found in Gillund and Shiffrin (1984). They explain the recognition advantage of low-frequency words over high-frequency words by assuming that when items are not rehearsed together, their residual inter-item strength depends on the frequency of the words. Low-frequency words have a lower inter-item strength than high-frequency words because high-frequency words are more often encoded together pre-experimentally than low-frequency words. This results in the low-frequency distractors having a lower strength than high-frequency distractors, thereby increasing the distance between target and lure strength distributions and producing a recognition advantage for low-frequency words. It is reasonable to assume that the same kind of process occurs for abstract and concrete words. However, even with these additional assumptions SAM has trouble explaining the structural effect. Because the inter-item strengths of lures depend only on the frequency and concreteness of the test items and not the study items, the inter-item strengths for lures would be the same regardless of the rule set used at study. That being the case, the familiarity associated with CC, RA, RC and CA words would be the same regardless of the rule set used at study. The predictions in such a case are that if a structural effect existed at all it would

exist for one rule set only whilst with the other rule set the structural effect would in fact be reversed, resulting in a negative structural effect – much like what was seen in the analysis of chunk strength. Thus it would appear that the basic SAM model does not predict the structural effect.

A second way that SAM could produce a structural effect is through how the model deals with similar lures. If a lure is similar to the list as a whole, this is reflected in the lure having a higher inter-item strength to all of the items on the list (Shiffrin et al., 1990). In the case of the structural effect, NC words would be similar to half the items on the list. Nothing in the SAM model specifies that conjunctions of concreteness and frequency produce similarity, but on the other hand this is not ruled out either. If SAM did pick up on the rule consistence as similarity it would be reflected in rule-consistent words having a higher base inter-item strength to study-list words than rule-inconsistent words. In this case, NC words would have higher familiarity than NI words. However, in SAM the repetitions also accentuate the effects of similarity and difference of lures to the study-items. In other words, at a low number of repetitions the effects of similarity are also low, but as the number of repetitions increase, so do the effects of similarity. So if SAM did somehow detect that NC words are conceptually similar to the study-list words, the difference in familiarity between NC and NI words would increase with repetitions resulting in an increase in the structural effect. So in conclusion, SAM is capable of predicting the existence of a structural effect provided that the model would count a conjunction as "similarity", but such a structural effect would increase by study strength.

5.3.1.2 MINERVA 2.

MINERVA 2 (Hintzman, 1984, 1986) is also a GMM model that has similar basic assumptions to SAM, except with MINERVA multiple images can exist of one word. MIERVA assumes two memory systems – long-term (or secondary) memory and the temporary working store (primary memory). The secondary memory consists of a number of traces. Each trace is a vector of features. Each feature can be either activated (takes a value of 1) inhibited (a value of -1) or irrelevant or not coded (a value of 0). For recognition memory, the secondary memory is assumed to consist of one trace for each item seen in the study phase. A learning parameter defines how well items seen at study are transferred into memory – the higher the learning parameter the fewer zeros the memory matrix will contain. The learning parameter also accounts for the effects of forgetting. Repetition of stimuli in a

study phase would lead to additional traces being created rather than existing traces being updated as occurs in SAM.

When an item is seen at test, a retrieval cue is sent by primary memory to secondary memory. The retrieval cue consists of the item with all its features coded with an additional set of context features that represent list membership. Secondary memory then responds with something called the echo. The echo depends on the match of the retrieval cue to all traces in memory. It contains two types of information – the overall intensity of the echo (a measure of the total match of the cue) and the content of the echo (the extent to which each individual feature of the cue matches the relevant features in memory). For recognition decisions the echo intensity is compared to a criterion. If the intensity is greater than the criterion, then an old decision is made, otherwise a new decision is made. One important feature of the process of matching the cue to memory is that a trace will only contribute significant echo intensity if there is a high degree of matching – low degrees of match do not contribute much to the echo intensity. The effect of repetition in MINERVA is simply to create new traces for each repetition. Thus a word repeated five times would have five traces all with slightly different features coded depending on the learning parameter (although all traces would be identical if the learning parameter were set to 1). Old words would thus have a higher echo-intensity with five rather than one repetition, as would new words. This produces problems for MINERVA in explaining strength-based mirror effects because the echo-intensity of old and new words are positively correlated. However, if the old word increase outstrips the new word increase, a strength-based mirror effect could be explained assuming a criterion shift also occurs.

The main problem for MINERVA and the structural effect is that it is not clear whether the rule-consistent words would have a greater degree of match to the study items than the rule-inconsistent words in terms of global concreteness and frequency. The study list contains the same number of concrete and abstract words and common and rare words, no matter what the rule set. However in MINERVA, a trace must have a high degree of match in order to contribute to echo intensity. Thus there are two possibilities. The additional degree of match due to the conjunction of concreteness and frequency could be negligible and NC and NI words produce the same echo intensity, because they equally match the global levels of concreteness and frequency on the study list. In this case, no structural effect would arise. Alternatively, NC words may produce enough of a match to rule-consistent words for NC words to produce higher echo intensity than NI words. In this case, MINERVA would produce a structural effect. If this were the case, it is likely (though not certain) that a study-

strength manipulation would increase the structural effect as the greater number of rule-consistent study traces would produce an even greater intensity for NC words. In order to test this theory, MINERVA simulations were conducted. For full details and discussion see section 5.3.4.

5.3.2 Likelihood models.

Likelihood-models are similar to GMMs in that they compare a probe to the contents of memory, but they differ in the basis of the decision. Two models are discussed – Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997) and Bind Cue Decide Model of Memory (BCD-MEM; Dennis & Humphreys, 2001). The Subjective Likelihood Model (SLiM; McClelland & Chappell, 1998) will not be discussed as although it is a distinct model from REM, in this case it makes similar predictions.

5.3.2.1 REM.

Similar to GMMs, REM assumes that memory consists of a series of images with each image containing a set of features. Each feature takes a value from zero upwards, with zero representing a lack of information about that feature. REM also assumes that the values occur such that one is the most common and the larger values are progressively more uncommon. The exact probability of each value occurring is a parameter of the model and is known as the environmental base-rate. When stimuli are memorised in a recognition memory study-phase, some features are coded appropriately, some are inappropriately coded and some are not coded at all. Inappropriately coded features take on a random value defined by the environmental base-rate. The proportion of features that are coded, both appropriately and inappropriately, are also parameters of the model.

When a stimulus is seen at test it is used to probe the images in memory. The features in the probe are compared with the features in each image in memory. For each image the pattern of matches and mismatches is computed, taking into account the environmental base-rate of features. Then, a likelihood ratio is computed for that image. The likelihood ratio is the probability of observing that pattern of matches if the image matches the probe divided by the probability of observing that pattern if the image is a word other than the probe. The likelihood ratios for each image are then combined into an odds ratio weighted by the number of images in memory. If the odds ratio is greater than one, an old decision is given. FAs occur because lures sometimes match images in memory by chance, because of inappropriately coded images or chance match of appropriate features. The central difference

between likelihood models and GMMs is that the likelihood approach allows for a process called differentiation. This means that not only are the similarities of a target probe to a target image taken into account, but also differences between lure probes and target images are taken into account. Consequently likelihood models can account for strength-based mirror effects without employing a criterion shift. In REM, a word seen in the study phase five times will have an image that contains more information than an image of a word seen only once. This increases the HR because the match of an "old" probe to a strong image will be better than that to a weak image. The FAR decreases because the probability of a chance match between a lure and the images in memory is lower in strong-study conditions than in weak-study conditions.

Factors such as word frequency are dealt with by varying the environmental base-rate of features for rare versus common words. This results in rare words having rarer features than common words, which results in rarer words having stronger matches to targets and weaker matches to lures. This effect is thought to be due to more than just differences in the frequency of occurrence of different letters in rare and common words (Malmberg, Steyvers, Stephens, & Shiffrin, 2002). It is reasonable to suppose that the same might be true of concrete and abstract words. If this is the case, then matches due to the rule set factors would be the same across the entire list for NC words and NI words resulting in no structural effect (because the study list contains the same number of common and rare words and the same number of concrete and abstract words).

It is possible that instead of the rule-set factors being individually considered, they are considered in some way that makes NC words more similar to target words than NI words. The mechanism for this in REM is not clear, but a study by Criss (2006) demonstrated that lures that were similar to individual target items elicited more FAs than lures that were not similar. The similarity in this case was that lures had one letter changed from targets, (e.g. boat and coat). In this case, the similarity operates because the words share many surface features, leading to matches occurring because many of the features encoded for targets are the same as features in similar lures. This mechanism only accounts for a structural effect if NC words share more features with the relevant study-list words than NI words do. Even then, the influence of study strength may be counteracted by the fact that any specific NC word would have a greater match to half the features of the study list, but would mismatch to a greater extent with the other half of the study list. In order for REM to properly account for the structural effect, one of the encoded features would need to be a feature indicating a conjunction of the concreteness and frequency. Criss found that the FA for similar lures

decreased more than for non-similar lures when subject to a study-strength manipulation. This indicates that if REM could account for the structural effect through specifically coding the conjunction, the structural effect would be sensitive to study strength. Thus REM fails at this hurdle – by predicting that the structural effect would be sensitive to study strength it fails to match the empirical data.

5.3.2.2 BCDMEM.

BCDMEM (Dennis & Humphreys, 2001) used a distributed set of nodes to represent items in memory. In a recognition study-phase a set of nodes is created to represent the study context, and these nodes become linked to the nodes representing the study items. A parameter (the sparsity parameter) defines how many of the possible study-context nodes are active at each viewing of an item and a learning parameter defines the probability of a link being created between the context and the viewed item. Additionally, items may also have links to the studied context due to pre-experimental factors (e.g. if a participant had been in a previous memory, experiment words on that list may have pre-existing links to a "studied-on-a-list" context-node). At test, a reinstated context is created for an item, with individual context nodes being activated depending on a forgetting parameter. This reinstated context is then matched to the context nodes that are activated by the test item. The number of matches and mismatches between these is then used to compute a likelihood ratio which is used to make the old/new decision.

An important characteristic of this likelihood ratio is that calculating it involves the participant making an estimate of the degree of learning that occurred at study. In terms of the model, this means that the likelihood ratio is calculated with an estimate of the learning parameter, and this estimate is a separate parameter from the *actual* learning parameter. This allows the model to take account of participant's beliefs about their own memory. This is the mechanism through which BCDMEM explains strength-based mirror effects – the better learning in the strong-study condition increases the HR, whilst the FAR drops because a change in the estimated learning parameter results in mismatches carrying more weight (Starns, White, & Ratcliff, 2010).

BCDMEM has some difficulty explaining effects related to FAs due to lures that are similar to study items. Dennis and Humphreys (2001) suggest three ways in which BCDMEM could explain such effects. The first is that when participants see a word in a study list, they implicitly generate other words in a similar category, and all of these

generated words are treated as being on the study list. This explanation would not lead to a structural effect, simply because it is highly unlikely to extend to such a higher order conjunctive category. For participants to implicitly generate rule-consistent words they would have to generate an inordinate number because the first words generated would presumably be those with a strong surface similarity (e.g. generating bat when studying cat) or those in a semantically related category (e.g. generating dog when studying cat). If a person implicitly generated words to the extent that other rule-consistent words are generated alongside the obvious words, the effective study list would consist of a truly huge number of words. This does not seem likely, and if it did occur FARs would be very high because of the sheer number of words that would register as old at test.

The second explanation is that participants notice either that a study list is categorised or that similar lures exist and change their criterion for these items. Participants in the experiments presented here did not notice the rule set at all.

The third explanation is that such effects are actually artificial and thus BCDMEM does not need to explain them. This explanation is uninformative in terms of the structural effect and so will not be discussed. BCDMEM has no other mechanism to create a structural effect. The pattern of matches and mismatches for lure items is not affected by the learning in the study phase. Matches and mismatches for lures only occur through pre-experimental familiarity. Since pre-experimental familiarity is the same for both NC and NI words, and similarity effects would not create a structural effect, BCDMEM itself does not predict the structural effect.

5.3.3 ACT-R.

ACT-R (Anderson et al., 1998) simulates memory with schema-like chunks that are linked together in a network. The linked chunks are used by production rules to define how and when the chunks are retrieved. When a particular chunk is activated, the activation spreads to any linked chunks with the amount of activation decreasing with an increase in the number of links. In recognition memory, a list chunk would be created which then links to the chunks representing the study-words and any linked attributes for those words. It is unlikely that the ACT-R model would predict a structural effect in the case of the experiments presented here. The activation from the concepts of concreteness and frequency would be the same for all the words on the list. Lures would create FAs due to activation spreading through these (and other) related concepts. Globally the study list creates exactly

the same base activation of all the concepts involved in the rule set in both conditions so no structural effect would be created.

5.3.4 MINERVA Simulations.

The predictions of MINERVA were tested with simulations. MINERVA has recently been used to test dual-process explanations of various effects. Several theories involving dual processes can in fact be explained by differences in parameters in MINERVA meaning that a single-process explanation can suffice (Jamieson et al., 2010; Jamieson & Mewhort, 2009a, 2009b). Similarly here, if a MINERVA simulation can produce the structural effect, then explanations in terms of multiple processes are not required.

MINERVA has a mechanism that attenuates the echo contribution of each instance in memory by the degree of match. In the discussion above it was unclear if this mechanism would result in a structural effect. Simulating the general pattern of data with MINERVA will provide an answer to this question. First, an expanded explanation of the MINERVA model will be provided.

5.3.4.1 The MINERVA model.

Long-term memory is represented by a memory matrix. Each line in the matrix (called a trace) represents an item. Each item is a set of features with values of 1, -1 or 0. When an item is entered into memory, each feature is either coded accurately with probability L, or not coded with probability 1-L. If an item is not coded a zero is entered into that position otherwise the actual feature is coded. Forgetting is simulated by setting different values of L.

To compare an item to memory, a probe is sent that consists of a perfectly coded copy of the current stimulus. MINERVA then returns an echo. The echo has two characteristics – the intensity and the content. Old/new recognition memory responses are made by comparing the echo intensity to a criterion. As the simulations are of the recognition memory case, only the echo intensity will be discussed here because the echo content is not used. The intensity depends on the similarity of each individual trace in memory to the test stimuli. The similarity (S) of trace i to the probe is given by Equation 1:

$$S(i) = (1 / N_R) \sum_{i=1}^{N} P(j)T(i, j) \quad (1)$$

P(j) is feature j of a probe, T(i, j) is the corresponding feature of memory trace i and N_R is the total number of features that are nonzero in the probe and the trace. Thus matches between the probe and the trace increase S(i), mismatches decrease it, and cases of 0 in either the probe or the trace are not considered. The extent of activation (A) of trace i is then given by Equation 2:

$$A(i) = S(i)^3 \quad (2)$$

The similarity is cubed, which has the effect of amplifying the contribution of traces that have a high match to the probe and suppressing the contribution of traces with a low match to the probe (i.e., nonlinear generalisation gradient). Finally, the intensity of the echo (I) is given by the summed total of the individual trace activations of all traces in memory (M):

$$I = \sum_{i=1}^{M} A(i) \quad (3)$$

A recognition decision is made on the basis of Equation 3 – intensities over a set criterion lead to old decisions whilst intensities below the criterion lead to new decisions.

5.3.4.2 Adapting MINERVA to the data.

In order to use MINERVA to simulate Experiment 5, the influence of pre-experimental frequency and concreteness must be amalgamated into the model. An initial set of simulations was conducted. The memory matrix for these simulations was prepared in the following fashion.

The words used in the experiment were generated by randomly producing a vector of 1s and -1s where P(1) = P(-1) = 0.5. The exact length of the vector was a parameter of the simulations. In the first set of simulations, concreteness was represented by an additional vector that was concatenated with each word vector. First, a vector was created to represent that the current word is abstract (again P(1) = P(-1) = 0.5). Each feature of this vector was then multiplied by -1 to create a vector for concrete words. This made concrete and abstract polar opposites of each other. Half of the words had the concrete vector added and the other half had the abstract vector added. The pre-existing memory-matrix was then created. Half of the concrete words and half of the abstract words were chosen to be common words and the other half of the concrete and abstract words were chosen to be rare words. Rare words were coded into memory once each (with a pre-experimental L depending on the simulation) and common words were coded into memory multiple times (the exact number was a parameter of the simulation). As each item was coded into memory a context vector was also added to represent the different contexts in which each word may have been encountered. The context vector was randomly created for each individual word. This resulted in common words having multiple error-prone traces in memory[6] and a variety of different contexts.

The study phase was then simulated by choosing 80 words that were rule consistent for the particular rule set. A study-list context vector was created which was appended to each of the word vectors. Each study item was then coded into memory with an experimental L. The experimental L was always larger than the pre-experimental L to simulate the effects of forgetting the pre-experimental items. This resulted in additional memory traces all with identical (error-prone) context vectors. Simulations were conducted in which study items were coded into memory once each and other simulations were conducted in which study items were coded into memory five times each in order to simulate the effects of study strength. In MINERVA, study strength results in new traces each time a word is encountered.

Each test word had the study-list vector concatenated onto it, and then this was used as a memory probe. The intensity of the echo for that word was recorded. For each simulated participant all of the vectors were randomised anew so that no two participants had the exact same set of vectors. For each set of parameters simulations were conducted for each rule set and at each level of study strength.

[6] This approach to frequency suggested by Randy Jamieson (Personal correspondence, 28th June 2010)

A second set of simulations was conducted using a slightly different method. A mirror effect is often found in which concrete words hold an advantage over abstract words (Glanzer & Adams, 1985). One way in which words can hold an advantage in recognition is if they were coded with different L parameters, perhaps because some words invite additional attention or are just easier to remember than others[7]. Thus in the second set of simulations, words did not have an abstract or concrete set of features but instead were coded both pre-experimentally and experimentally with different L parameters.

The resulting data were not compared directly to the experimental data. Rather, the pattern of the echos was analysed. If MINERVA can simulate the general pattern from Experiments 3, 4 and 5 then several things should be the case.

- The intensity of the echo for NC words should be greater than the intensity for the NI words.
- This should be the case regardless of the rule set used (i.e. CCRA or RCCA).
- Finally, if the NC intensities are greater than the NI intensities, the magnitude of the difference should not be affected by the study-strength manipulation.

In order to test for this pattern, the intensities from each simulation were entered into an ANOVA with study strength (weak-study versus strong-study), rule set (CCRA versus RCCA) and word type (either old versus NC or NC versus NI). One ANOVA was run for the episodic effect (old versus NC) and another for the structural effect (NC versus NI).

5.3.4.3 Results.

Table 5.5 shows the different parameters tested. Note that simulations where the abstract and concrete L values are identical were from Simulation Set 1 in which concreteness was represented with vectors. Simulations with different abstract and concrete Ls were from Set 2, and thus do not have a value for the length of vector used to represented concreteness.

[7] Different L parameters for concrete and abstract words suggested by Hintzman (personal correspondence, 24th August 2010)

Table 5.5

Parameters for MINERVA Simulations

Simulation	Set	L_{pa}	L_{pc}	L_{ea}	L_{ec}	F_w	F_{co}	F_c	Common	Participants
1	1	.20	.20	.55	.55	30	10	20	100	400
2	1	.20	.20	.55	.55	30	10	20	50	400
3	1	.20	.20	.55	.55	30	10	20	50	200
4	1	.40	.40	.55	.55	30	10	20	50	200
5	1	.20	.20	.55	.55	30	10	10	50	200
6	1	.20	.20	.20	.20	30	10	20	50	200
7	1	1	1	1	1	30	10	20	50	200
8	1	.20	.20	1	1	30	10	20	50	200
9	1	.20	.20	.55	.55	30	4	20	50	200
10	2	.20	.30	.55	.65	30	-	20	50	200
11	2	.20	.25	.55	.60	30	-	20	50	200
12	2	.20	.30	.55	.65	30	-	20	100	200
13	2	.45	.50	.75	.80	30	-	20	50	200
14	2	.20	.30	.55	.65	30	-	10	50	200
15	2	.20	.30	.55	.65	20	-	10	50	200

Note. Set = Simulation set (concreteness represented as vectors in set 1 and as different Ls in set 2), L_{pa} = Pre-experimental abstract learning parameter, L_{pc} = pre-experimental concrete learning parameter, L_{ea} = experimental abstract learning parameter, L_{ec} = experimental concrete learning parameter, F_w = number of features in word vector, F_{co} = number of features in concreteness vector, F_c = number of features in context vector, common = number of times common words replicated in memory, participants = number of simulated participants.

The patterns of data and results for Simulations 1 through 9 were identical as were the data for Simulations 10 through 15. Thus the analyses for simulations 3 and 10 were randomly chosen to be presented. Simulation 3 is wholly representative of the results for simulations 1 through 9 and simulation 10 is wholly representative of the results for simulations 10 through 15.

5.3.4.3.1 Simulation 3.

For Simulation 3 there were no effects of rule set so the data was collapsed across this factor. See Table 5.6 for means and standard errors.

Table 5.6

Mean Echo Intensities by Word Type and Study Strength for Simulation 3 (SE in brackets)

Word-type	Weak	Strong
Old	1.39 (.01)	6.57 (.01)
NC	1.25 (.01)	5.86 (.01)
NI	1.25 (.01)	5.84 (.01)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

The ANOVA with old and NC word types had a main effect of word type $F(1, 196) = 4650.71$, $p < .001$, $\eta^2 = .96$, a main effect of study strength $F(1, 196) = 105{,}727.31$, $p < .001$, $\eta^2 = 1.00$ and an interaction between study strength and word type $F(1,196) = 2089.60$, $p < .001$, $\eta^2 = .91$. There were no other effects, all $Fs < 1$. The word-type effect indicated higher intensity for old words ($M = 3.98$, $SE = .01$) than for NC words ($M = 2.55$, $SE = .01$). The study-strength effect indicated higher intensity with strong-study conditions ($M = 6.21$, $SE = .01$) than with weak-study conditions ($M = 1.32$, $SE = .01$). The interaction indicated a greater difference between old and NC intensities with strong-study conditions, $F(1, 198) = 84{,}154.93$, $p < .001$, $\eta^2 = 1.00$, than with weak-study conditions, $F(1, 198) = 96{,}111.18$, $p < .001$, $\eta^2 = 1.00$.

The ANOVA with NC and NI word types had only a main effect of study strength, $F(1, 196) = 96{,}165.04$, $p < .001$, $\eta^2 = 1.00$. This indicated higher intensity in strong ($M = 5.85$, $SE = .01$) rather than weak ($M = 1.25$, $SE = .01$) study conditions. There were no other effects, highest $F(1, 196) = 2.33$, $p = .13$.

5.3.4.3.2 Simulation 10.

Unlike with Simulation 3, Simulation 10 had an effect of rule set. The simulation 10 data was therefore not collapsed across rule set – see Table 5.7 for means and standard errors.

Table 5.7

Mean Echo Intensities by Rule Set, Word Type and Study Strength for Simulation 10 (SE in brackets)

Word type and rule set	Weak	Strong
CCRA		
Old	2.07 (.02)	9.51 (.03)
NC	1.87 (.02)	8.50 (.03)
NI	1.73 (.02)	8.36 (.03)
RCCA		
Old	1.96 (.02)	9.33 (.03)
NC	1.76 (.02)	8.29 (.03)
NI	1.89 (.02)	8.44 (.03)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words, CCRA = common-concrete/rare-abstract, RCCA = rare-concrete/common-abstract.

The ANOVA for old and NC words had main effects of word type $F(1,196) = 4638.89$, $p < .001$, $\eta^2 = .96$, study strength $F(1, 196) = 80{,}727.04$, $p < .001$, $\eta^2 = 1.00$, rule set $F(1, 196) = 4885.37$, $p < .001$, $\eta^2 = .16$ and a study-strength by word-type interaction $F(1, 196) = 2099.55$, $p < .001$, $\eta^2 = .91$. There were no other effects, all $Fs < 1$. The word-type effect indicated greater intensity for old words ($M = 5.72$, $SE = .01$) than for NC words ($M = 5.10$, $SE = .01$). The study-strength effect indicated greater intensity in strong-study ($M = 8.91$, $SE = .01$) than in weak-study conditions ($M = 1.92$, $SE = .02$). The rule-set effect indicated greater intensity for CCRA words ($M = 5.49$, $SE = .02$) than for RCCA words ($M = 5.34$, $SE = .02$). The interaction indicated a greater difference between old and NC word intensities with strong-study, $F(1, 196) = 6490.05$, $p < .001$, $\eta^2 = .97$, rather than weak-study conditions, $F(1, 196) = 248.39$, $p < .001$, $\eta^2 = .56$.

The ANOVA for NC and NI words had a study-strength main effect, $F(1, 196) = 81{,}356.15$, $p < .001$, $\eta^2 = 1.00$, and a word-type by rule-set interaction $F(1, 196) = 183.53$, $p <$

.001, η^2 = .48. The effect of study strength indicated greater intensities with strong-study (M = 8.40, SE = .02) than with weak-study conditions (M = 1.81, SE = .02). The rule-set by word-type interaction indicated that with a CCRA rule-set, NC words had higher intensities than NI words, $F(1, 196) = 89.17$, $p < .001$, $\eta^2 = .31$. With a RCCA rule-set, NI words had higher intensities than NC words, $F(1, 196) = 94.39$, $p < .001$, $\eta^2 = .31$. There were no other effects, all $F < 1$.

5.3.4.4 Discussion.

The results from Simulation 3 (and from simulations between 1 and 9) in which concreteness was instantiated with a vector were fairly straightforward. MINERVA did not produce the structural effect. The echo intensity for old words was higher than that for NC words and this difference increased with study strength. Thus simulation 3 successfully simulated the pattern of data for the episodic effect insofar as the pattern of changes in the echo intensity would lead to the experimental data if participants used the intensity of the echo to make their decisions.

The results of Simulations 10 onwards in which concreteness was instantiated by different learning parameters for concrete and abstract words demonstrated that NC intensities would be higher than NI intensities for the CCRA rule-set. For the RCCA rule-set on the other hand, NI intensities were higher than NC intensities. Far from being a structural effect, this simply indicated that the CCRA words resulted in higher intensities than the RCCA words regardless of the study list used. In fact, collapsing across rule set in this case would eliminate the structural effect and yield results similar to Simulations 1 through 9.

There are several potential criticisms that could be levelled at these simulations. Although concrete and abstract words were instantiated in different ways, word frequency was instantiated in the same way in each simulation. An alternative approach would have been to apply different L values to common and rare words as suggested by Hintzman (1988). This was not done for two reasons. The first is that MINERVA is a multiple-trace model and so representing common and rare words as words that have been encountered a different number of times is intuitive both by the model logic and by ecological validity. Hintzman argues that common and rare words differ in the number of salient features they each contain, but if rare features are salient because of the fact that they are rarely seen, it would be more accurate to say that the number of salient features is a correlate of word frequency and not the defining factor. What makes a word common or rare is by definition the frequency with which the words are encountered. The second reason is that, as can be seen in the difference

between the first and second set of simulations, applying different L values to different word types results in a fixed ordering regardless of what occurred in the study phase. Thus this change would be unlikely to result in a structural effect being produced by the model, although it may change the ordering of the word types.

A second criticism is that an alternative way to analyse the model data would be to set different criteria and match the pattern of hits and FAs to the experimental data. It is possible that for some patterns of echo intensities, setting a particular criterion might result in a structural effect. For instance, if the NC words had echo intensities of 1, 1, 3 and 4 whilst the NI words had echo intensities of 1, 2, 2 and 4, most criteria would result in no structural effect (and the mean echo intensities for NC and NI words would be the same). A criterion set between 2 and 3 would result in 2 NC words being called old and 1 NI word being called old, resulting in a structural effect. Although setting a small set of particular criteria may create a structural effect, it is certain that setting other criteria (in fact the majority of criteria) would produce no structural effect as can be seen by the general pattern of the intensities from the simulations. Thus the overall characteristics of the model do not predict the structural effect. This is a point well made by Cleeremans and Dienes (2008). They point out that a model must fit the way that humans work rather than fitting model outputs to data in order to explain human performance. In other words, the way that a model works must be by the fit of its general characteristics to the pattern of human data. If changing various parameters results in the model fitting or not fitting the data, then the model is not a model of human behaviour, it is just a model that can produce a particular set of numbers. MINERVA by its general characteristics appears to have trouble dealing with a conjunctive rule-set because it has no mechanism for representing such a conjunction. Therefore, it does not adequately model the pattern of performance obtained in the experiments presented in the preceding chapters.

In general then the recognition-memory models discussed here appear to have a problem dealing with the conjunctive nature of the rule set. Partly this is due to the fact that most of them have no means of representing the conceptual nature of concreteness. Concreteness does not rely in any way upon the surface characteristics of the words, which means that methods of coding chunks do not do the job. Instead, models need two characteristics to deal with the structural effect. They need a way to represent meaning and conceptual features and they need a way to represent conjunctions across both conceptual and surface features. This is a difficult hurdle for several reasons. Once a model starts to code for conjunctions the number of conjunctions becomes infinite. There could be conjunctions

between any number of conceptual features and many conjunctions would occur by chance. Additionally there could be conjunctions of conjunctions, or conjunctions of conjunctions of conjunctions, resulting in an infinite number of possible conjunctions to track. Any model including a conjunctive mechanism would have to deal with that complexity. Furthermore, it is not clear how concepts might be represented in these models. To use the example of MINERVA, a concept would be represented by a combination of particular feature settings. Each feature represents a "primitive property" such as colour or odour. But then what combination of these features creates the concept of concreteness? Is concreteness itself a primitive property that can be represented by one feature alone? Or is it a combination of properties such as "can be touched" and "can be seen"? Clearly defining the vectors as in the current simulations does not represent them adequately. It is not conclusive that MINERVA cannot adequately represent concreteness, but what modifications might allow MINERVA to do so are not clear.

Overall, none of the models of recognition memory discussed here predict a structural effect, or predict invariance of a structural effect to study-strength manipulations. In order for these models to be complete, they must be able to explain phenomena such as implicit learning of rule sets that rely on more than just surface characteristics of stimuli. In this sense, a model that explains AG performance is not sufficient to explain implicit learning as a whole.

6 Chapter 6: General Discussion and Final Conclusions

This final chapter presents a summary of the findings from the previous chapters before discussing the implications and limitations of the research.

6.1 Summary of Findings

Chapter 1 reviewed concepts from implicit learning and recognition memory and concluded that the methods used in recognition memory could help shed light on implicit learning.

Chapter 2 employed stimuli consisting of natural words that were classified according to a conjunctive rule-set based on concreteness and frequency. It was shown that participants learnt the rule set and could discriminate rule-consistent words from rule-inconsistent words. Questionnaire measures indicated that they did so without verbalisable knowledge of the rule set. Multiple subjective measures of awareness varied in their results. Participants' discrimination performance was unaffected by distraction at study or test for rule-consistent words. Performance for old words was reduced by distraction at test in the memory category only. These results demonstrated that a rule set instantiated in natural words could be learned and used even though no verbalisable knowledge of the rule set could be expressed. Natural words thus provided a stimuli set that could be used in both recognition and classification tasks allowing links to be drawn between recognition memory and implicit learning.

Chapter 3 employed these stimuli in a set of experiments that involved both recognition task and classification tasks. Using SDT, the influence of rule-based performance (the structural effect) was distinguished from the influence of memory for individual words (the episodic effect). Structural knowledge influenced participants' performance regardless of whether it was appropriate, as in a classification task, or inappropriate, as in a recognition task. The structural effect was shown to be insensitive to manipulations of study strength and depth of processing, whereas the episodic effect was sensitive to both manipulations. The strength-based mirror effect was shown to occur in both classification and recognition tasks with increases in study strength resulting in increases in HRs and decreases in FARs. The task demands were also shown to affect participants' performance through criterion shifts. Classification instructions resulted in participants adopting a more liberal criterion but reducing reliance on the episodic status of words compared to recognition instructions. Although the structural effect itself was insensitive to study strength, participants' metacognitive states did seem to be affected. Confidence for correct answers to NC and NI words was higher than confidence to incorrect answers in the

strong-study condition but not in the weak-study condition. The attribution data suggested that study strength mainly affected endorsements in the memory category. These results demonstrated that the processes underlying recognition and classification tasks are similar. Even though familiarity has been implicated as a factor in both recognition and classification tasks previously, the familiarity underlying the structural effect seemed to be different from that underlying the episodic effect.

Chapter 3 also demonstrated that an instructional manipulation did not change the magnitude of the structural effect, although there were some indications that instructions to look for commonalities between words at study could interfere with the structural effect. A PLUM analysis indicated that the NC and NI item distributions had similar variance whilst the old distribution had larger variance than the NC and NI distributions. The NC distribution looked more like a lure distribution even though in classification it is a target distribution. The traditional signal-detection measure *d'* was thus shown to be appropriate for measuring the structural effect.

Chapter 4 demonstrated that the structural effect was not affected by deadline and distraction manipulations. The episodic effect still increased by study strength under short-deadline conditions, although not in the classification task. Participants appeared to shift their criterion on a trial-by-trial basis in response to the deadline manipulation. A distraction manipulation had little effect on the episodic effect. Distraction did appear to prevent participants from shifting their criterion between classification and recognition. The evidence from the deadline experiment provided further evidence that the familiarity underlying the structural effect is different from the familiarity underlying the episodic effect.

Chapter 5 addressed some possible methodological limitations. Chunk strength was dismissed as the basis of the structural effect. It was shown that even when the participants showing the strongest structural effects are isolated, the structural effect was still insensitive to study strength. Collapsing across several experiments demonstrated that this insensitivity was not a problem with statistical power. Finally, a review of memory models revealed that current models do not predict a structural effect that is insensitive to study strength. Several MINERVA simulations confirmed this, as MINERVA predicted the same ordering of stimuli types regardless of rule set used at study. Thus Chapter 5 provided further evidence that the familiarity underlying the structural effect is not sensitive to study strength, and that existing explanations of implicit learning and recognition memory do not predict this result.

In summary, the experiments presented here have demonstrated that methods used in recognition memory and implicit learning can be fruitfully combined. Doing so has revealed

that both structural and episodic effects contribute to both recognition and classification tasks, and that the contribution of the structural effect is not sensitive to study strength. These results have implications for both implicit learning and recognition memory, which will now be discussed.

6.2 Implications

Chapters 1 and 2 posed the following questions:

- What is learned in implicit learning?
- In what areas can implicit learning and recognition literatures develop more cross talk?

This section will address the first of these questions by discussing the implications of the experimental results for recognition memory and implicit learning. The second question will then be addressed by discussing the results in terms of a combination of the recognition memory and implicit learning literatures. Possible reasons as to why the structural effect was not sensitive to strength are also discussed.

6.2.1 Implications for recognition memory.

As in Higham and Brooks' (1997) experiments, the recognition task displayed influences of both the structural status of the word (the structural effect) and the episodic status of the word (the episodic effect). The fact that the structural effect was not sensitive to study strength builds on Higham and Brooks' results. A structural effect that is not sensitive to study strength falsifies the recognition-memory models reviewed in Chapter 5 because none of them predict the structural effect's existence and response to study strength. Such falsification is at least one way in which the experiments presented here can be said to be a scientific contribution (see Dienes, 2008b for discussion of different views of Psychology as a science).

As discussed in Chapter 5, recognition-memory models do not produce the pattern of predictions necessary for a strength-insensitive structural effect to exist. Broadly speaking, the models represent stimuli in one of two ways. REM (Shiffrin & Steyvers, 1997), MINERVA (Hintzman, 1984, 1986) and (at a slightly abstract level) SAM (Gillund & Shiffrin, 1984) all represent stimuli with a vector of features. BCDMEM (Dennis & Humphreys, 2001) and ACT-R (Anderson et al., 1998) both represent stimuli as nodes in a network. There are two central problems that these representative structures face when trying to predict the structural effect. The main problem is that none of these ways of representing

stimuli give any special status to conjunctions required to produce the structural effect. Without a separate representation of conjunctions any model that makes an old/new recognition decision by globally comparing NC and NI words to the study list will fail because the study list has a balanced number of rare, common, abstract and concrete words. If a model did predict a structural effect the problem then becomes that most models predict that a study strength increase will create some kind of change in a structural effect. In MINERVA for instance if a stimulus has a greater echo than another an increase in study strength would increase this difference. Models that use differentiation such as REM (and ALT (Glanzer & Adams, 1990)) predict changes in the structural effect because increase study strength results in increased differentiation, and similar lures are affected more by differentiation than are non-similar lures. Since some form of similarity is the only mechanism differentiation models have of predicting the structural effect, such a structural effect should be sensitive to study strength.

Changes clearly need to be made to the models to allow them to produce a structural effect of the sort identified here. The model that would need the smallest amount of change in this regard is SAM. In the original application of the SAM model to recognition memory (Gillund & Shiffrin, 1984) the strength of lures did not depend on the study list, only on pre-experimental factors, reflected in the fact that any word had a certain base activation to any other word from previous occasions when they had been processed together. Later on, the base activation of lures was allowed to vary with the degree of similarity of the lure to all or part of the study list, reflected in similar lures engendering higher base activation rates from study list items than dissimilar lures. Additionally, a differentiation mechanism reduced the influence of lure similarity as study strength increased (Shiffrin et al., 1990). If the differentiation mechanism was removed from the model but the lure similarity aspect retained, SAM would come close to predicting the structural effect. SAM does not specify how exactly the model rates the similarity of a lure to the study list. If conjunctions of features were somehow rated as more "similar" than individual features were, then SAM can produce the observed pattern of data. NC words would elicit more old responses than NI words because they were more similar to the study list, and this difference would not be increased by study strength because the strength of lures would not increase with study strength. Of course this would prove problematic for SAM overall because the differentiation mechanism was introduced to explain phenomena such as the strength-based mirror effect. It is also important to note that SAM can only simulate the data because similarity is ill-defined in the model, and similarity tends to refer to surface similarities, category membership and

obvious semantic relations. Thus if similarity were well defined in SAM, it is likely that such a definition would not rate a conjunctive rule set as a factor that produces similarity.

In terms of the other models, some mechanism for dealing with conjunctions would need to be introduced for them to produce a structural effect. This would greatly increase the amount of information they handle and the complexity of the models. Most of the memory models restrict themselves to a small number of simple mechanisms and attempt to keep the computational burden low. These are both perfectly reasonable aims when creating an approximate model for how a specific process might work, but they may not be founded when making claims about the processing power involved in human behaviour. The computational ability available to humans is not currently known, but it is reasonable to assume it must be quite substantial. Taking actions such as walking without falling over or losing balance requires complex calculations. We do this in real time with no appreciable effort and plenty of capacity left to enable us to talk, listen to music and chew gum without biting ourselves all at the same time. In this sense, the computer metaphor may be misleading because it encourages researchers to compare the processing capabilities of humans to the current capabilities of computers. However, unless we operate in the same way as computers the comparison is an empty one. Conjunctions pose a problem not because they require *some* additional computational ability, but because the computational ability required to process them could in fact be infinite. Once conjunctions of individual features are tracked, why not track conjunctions of conjunctions? This could be extended to deeper and deeper levels. Such a mechanism would have to be limited in some way, perhaps to only tracking first order conjunctions (i.e. conjunctions of actual features rather than conjunctions of conjunctions). This would still allow for a great deal of complexity but would not require infinite processing resources. The data presented here only supports such first-order conjunctive learning, but it would be of interest to test the limits of such conjunctive learning.

Even with a method for tracking conjunctions, computational models would struggle with the fact that the study-strength manipulation did not increase the structural effect. This could be accounted for if the parameters that govern learning in computational models (such as L in MINERVA) do not increase in a linear way with study strength. One possibility is that the learning parameter is significantly reduced after the first repetition. In a model like MINERVA this would still result in an increase in the echo intensity between NC and NI words, but if the L was small enough the additional difference between the NC and NI echo intensities could be negligible, especially if the conjunction is embodied in a small number of features. A second possibility is that the learning parameter is not the same for all features

across one vector for repeated stimuli. In MINVERA, the first time a stimulus is encountered all features could be coded with the same L (L1). The second time a stimulus is encountered, a new instance is created for that word but some of the features are coded with L1 and some of the features are coded with a reduced L (L2). The reduction in L from L1 to L2 for these features could reflect a saving in processing power for repeated stimuli. In this way, processing of certain commonly used simple features could be prioritised in repeated stimuli, and the processing of more complicated and infrequently used conjunctions could be suppressed. With no additional coding of the conjunction, the effect of the repeated stimuli would be to increase the overall echo intensity, but not the structural effect, because the absolute gap between the NC and NI echo intensities would be the same regardless of the number of repetitions. Thus the structural effect would not increase with study strength.

Such a differential processing explanation could explain differences in the results presented here and one study using REM (Criss, 2006). Criss found that lures similar to individual study stimuli produced more FAs than normal lures, producing an effect similar to the structural effect. However, study strength decreased FAs to similar lures to a greater extent that FAs to normal lures and the magnitude of this difference was sensitive to the degree of similarity of the similar lures. For very similar lures, FAs can even increase with study strength. The sensitivity of the similar lures to study strength suggests that different factors underpin Criss's results than those that underpin the structural effect. The similarity of the stimuli hinged on surface-level similarity whereas the studies presented here hinged on a complex conceptual similarity which may operate in an entirely different way. This difference could be explained if the processing of different aspects of the stimuli are selectively stopped at different times as suggested above. Factors that drive surface similarity such as the letters in the word could continue to be analysed up to multiple repetitions of a word whilst the analysis of conjunctions of conceptual factors may be halted based on the continued presence of these conjunctions in the new stimuli. In order to prevent wastage of resources, processing of the conjunctive rule-set is stopped after just a small number of stimuli that have that conjunction.

Most of the computational models of memory imply a single-process approach. Other single-process explanations might suggest that the structural effect is based on weak episodic memory that is not verbalisable (Jimenez et al., 2006). In this case, the knowledge should become more verbalisable with an increase in study strength. This was not supported in the results of the experiments from the previous chapters. Instead at least a dual-process explanation is required to explain the results in which the structural effect is driven by one

process that is not sensitive to study strength and the episodic effect is driven by another process that is sensitive to study strength.

One of the most common assertions in recognition memory literature is that recognition memory is driven by recollection and familiarity (Yonelinas, 2002). Familiarity in recognition memory appears to be sensitive to study strength (Jacoby, 1999). Since the structural effect does not behave like familiarity does in recognition experiments, it would appear to be driven by different factors than recognition-memory familiarity. It is hard to say if such 'structural familiarity' is experienced any differently than episodic familiarity by participants. The results did not provide enough evidence to distinguish between an explanation involving one feeling of familiarity with two different components or an explanation involving two different feelings of familiarity. It is possible that differing results attributed to familiarity in recognition memory literature could reflect differences in which type of familiarity (episodic or structural) was being used.

It is easy to see why a structural effect would exist in terms of recognition memory. Such information allows us to distinguish between complex classes of object or people without having to reference conscious knowledge of the features which drive the categorisations. General categories of items are good examples of this – when I see a table I know it is a table without having to consciously think about all the attributes of a table, or consciously recall every instance of a table I have seen previously. The world is full of such conjunctions that are useful for us to remember.

6.2.2 Implications for implicit learning.

Most implicit-learning models do not account for the results of the experiments presented in the previous chapters. As demonstrated in Chapter 2, participants used several different factors to help them make their decisions in a classification task. Participants were sensitive to both the episodic status of words and to the structural status of the words. The structural status of words was accompanied by no verbalisable knowledge.

A.S. Reber's (1967, 1989) account of implicit learning hinges around participants abstracting knowledge of the underlying rules of the stimuli. This account could explain the existence of the structural effect provided that some mechanism for learning conjunctions is assumed. For such a structural effect, an increase in study strength could result in better learning of the rule set as increased exposure to the stimuli allowed more opportunity for the rules-learning process to operate. This would shift the NC distribution to the right from weak-study to strong-study conditions. However, also consistent with A.S. Reber's account

would be that in the strong-study condition there is a greater attempt to learn rules explicitly, which would interfere with the implicit rule-learning process and maybe even result in worse performance in the strong-study rather than weak-study conditions. This would result in the NC distribution shifting left from weak-study to strong-study conditions. Either interpretation implies a change in the structural effect by study strength. It could be possible that both influences occur and effectively cancel each other out.

A chorus-of-instances theory (Vokey & Brooks, 1992; Vokey & Higham, 2005) would require a similar cancelling of influences to explain the invariance of the structural effect to study strength. Viewing a stimulus at test triggers the retrieval of all study stimuli that are similar to it. The more instances are retrieved, the greater the strength-of-evidence felt. Vokey and Brooks investigated the effects of individuating stimuli for this chorus-of-instances theory. A stimulus is individuated when minor differences are emphasised and commonalities are suppressed. In theory, this would result in items not seeming similar enough to form a pool of items to be used in a chorus-of-instance type process. In several experiments Vokey and Brooks found that individuating items reduced the contribution of similarity to performance but left grammaticality influences untouched. They concluded that both chorus-of-instance and rule-learning processes existed and were not necessarily linked under the same control mechanism. If the structural effect was based on a chorus-of-instance type model (MINERVA being an example), then individuating the stimuli would reduce its magnitude. It is possible that study strength in some way individuates the stimuli, while simultaneously allowing a separate rule-learning system to better learn the rule set. Thus as for A.S. Reber's (1967, 1989) account, the two influences could cancel each other out resulting in no change in the magnitude of the structural effect by study strength. Such balance seems unlikely because the two influences involved for each theory would have to perfectly balance each other out for three repetitions of study words, five repetitions of study words, under deep and shallow encoding conditions and under different display times of the study-stimuli. All of these factors changed in Experiments 3 through 5. In order to truly rule out this possibility, manipulations would have to be employed that differentially alter the contributions of each influence. If the two influences are balanced, throwing them out of balance should result in a structural effect that is sensitive to study strength in some way.

Various studies have demonstrated that performance in AG tasks can be explained by surface features such as constraints on letters in the strings (Jamieson & Mewhort, 2009a) and letter repetitions (Brooks & Vokey, 1991; Vokey & Higham, 2005). The structural effect is more in line with studies that indicate that surface structures can be used but that other

processes also operate (Higham, 1997a; Tunney & Altmann, 2001). For reasons discussed above, any explanation hinging on surface features would be sensitive to study strength. Given that surface structure is learned in some instances, suggestions of flexible systems that learn different factors depending on the situation (Whittlesea & Dorken, 1993) seem plausible. Such a flexible system may produce a structural effect that is resistant to changes in study strength because of how the system switches between learning modes. It could be that such a system learns about the rule set and chunks and that study strength increases the degree of learning of the chunks but not of the rule set. This system would be capable of explaining the fact that the episodic effect increased by study strength – the increased chunk learning would lead to an increased episodic effect whilst the static learning of the rule set would produce the same structural effect in weak-study and strong-study conditions.

The structural effect provides evidence that participants can learn a conjunctive rule-set that crosses a conceptual factor (concreteness) with a statistical factor (frequency). Most implicit-learning experiments focus on purely statistical factors. SRT experiments involve learning relationships between stimuli positions and AG experiments are about which letters occur after which other letters. Neither of these paradigms require the use of conceptual factors. In a sense, learning AGs or SRT sequences is unambiguous – there is always a right and a wrong answer. Even when the stimuli have relationships involving some degree of variation, given enough stimuli a specific unambiguous relationship can be divined (e.g. a B follows an A 60% of the time but the other 40% of the time a C follows an A). Having concreteness in the conjunctive rule-set demonstrates that learning can occur in more ambiguous circumstances. It is reasonable to question whether the conjunctive rule-set used was at all ambiguous – after all, it can be defined in a definite way. However, where concrete words are concerned there is always the possibility of ambiguity. Take as an example the word table, presented out of context. This could bring to mind a specific table, or the table I am sitting at now. In this sense the word is definitely concrete. However, it could also refer to tables in general, or the overall category of tables. In other words the word table could be representing the property of "tableness" rather than a specific table. This sense of the word is abstract – the category of "table" cannot be touched. To further complicate things, table can also be a verb, as in to table a motion, which would be classified as abstract. Many concrete words have this ambiguity about them – they can be concrete or abstract depending on context. This could be one reason why the structural effect is small – because the degree of learning is dampened by the ambiguity of the relationship between concreteness and frequency.

Earlier in this chapter, I suggested that changes in how the stimuli are processed could limit the sensitivity of the structural effect to study strength. Some implicit learning evidence supports this theory. Rausei, Makovski and Jiang (2007) conducted an experiment examining how attention interacted with a visual-search task. They found that a small amount of attention was needed to perform the task, but above that extra attention did not increase performance. This is similar to the ideas suggested in the recognition-memory section above – a small amount of learning could result in a structural effect, but extra learning does not increase it beyond that. Hintzman and Curran (1995) provided further evidence for differential processing with their registration without learning phenomenon. Participants were shown words in a study phase, with some of the words being shown multiple times up to a maximum of 20. Participants were then asked to make frequency judgements at test. Similar lures were used in that some test words were plurals of singular words shown at study or singulars of plural words shown at study. At low numbers of repetitions (up to three) participants learnt to discriminate between similar lures and actual targets. However, at more than three repetitions, participants did not develop any greater level of discrimination, as similar lure FARs increased in proportion to the HR. Hintzman and Curran concluded that wastage of processing resources is prevented by only analysing novel stimuli fully. Once stimuli or a pattern becomes expected, only frequency of appearance information is further noted, which would support the increase of the similar lure FAR. Again, the conjunctive rule-set could quickly become expected and thus not processed. In essence, this is an efficiency explanation – ignoring some information means more salient-seeming information can be prioritised. If this is the case, then the structural effect may be limited to a particular magnitude (between 0.25 and 0.35 perhaps) and below that level study strength may increase or stabilise the structural effect until it reaches that limit.

From an evolutionary perspective, the structural effect can be thought of in terms of A.S. Reber's (1989) ideas of primitive and complex systems. Reber suggested that there are primitive systems that deal with basic information that need no prior input, such as counting and covariance of simple environmental features, and complex systems that perform processing involving either the outputs of the primitive systems or else more meaning-based processing. Learning a conjunctive rule-set involving the concept of concreteness would certainly involve A.S. Reber's complex systems. Ages back when survival would have been the priority, conjunctions involving abstract features would have been needed in order to survive. At the simplest level the conjunction of the physical feature of "sharp teeth" and "death" would have needed to be learned quite quickly. The same is true of conjunctions

such as “red”, “round” and “poisonous” which would have aided choices in berry foraging. In other words, development of these complex systems would have resulted in greater survival rates for our ancestors.

6.2.3 Synthesising the recognition memory and implicit learning perspectives.

The fact that the recognition and classification tasks produced similar data demonstrates that there should be closer links between recognition and implicit-learning literatures. A link between classification and recognition is not a new concept – for instance, Higham has pointed out the fact that both tasks seem to be performed using the same processes (Higham, 1997a; Higham & Brooks, 1997). In fact, when talking about implicit learning theories, it would seem to be difficult not to talk about recognition memory at the same time. The chorus-of-instances explanation requires memory for each instance. Learning of abstract rules requires remembering the rules. Global-memory models require noting individual features of stimuli which could result in the learning of some rule-based information. The only real difference in the data between recognition and classification tasks appears to be the reduced episodic effect in classification. This is likely to be due to the classification instructions encouraging less use of recollective information than is used in the recognition task. In essence, there is no reason to talk about learning and memory separately in different literatures. Learning cannot occur without memory and one important use of memory is to support learning. Essentially, all recognition-memory models and all implicit-learning models should be one and the same thing. In this way, the contributions of each literature can be leveraged to the greatest extent.

The fact that similar processes drive classification and recognition tasks require that care be utilised when talking about either task. Familiarity is implicated as a process that drives episodic responses and also structural responses. This still leaves two possibilities as to how this occurs. There could be just a single familiarity process to which many different sources contribute, resulting in a feeling of familiarity which does not change depending on the exact content of that feeling. Alternatively, it could be that there are different kinds of familiarity which are open to introspection and thus distinguishable from each other. Either possibility requires researchers to be specific about what kind of familiarity they are discussing.

One difference between the classification and recognition data is that there were higher FARs in the classification data, whilst the structural effect remained unaltered. This pattern of results is easy to explain with a criterion shift explanation of the strength-based

mirror effect (Stretch & Wixted, 1998). Participants simply adopted a more liberal criterion in classification in response to the difficulty of the task. The reduction in old word endorsements can be explained by participants using a reduced amount of recollection in classification than in recognition. This might happen because the classification instructions do not specify that one way to classify words as rule consistent is to look out for old words and thus perhaps participants missed some old words in classification. Differentiation models such as REM (Shiffrin & Steyvers, 1997) have more difficulty in explaining this pattern of results. Nothing has changed about the number of instances that have been stored, or the learning rates at which they were stored. Thus the pattern of matches and mismatches for any one stimulus should be the same in classification and recognition. Additionally, although there is a criterion in models such as REM, such a criterion is usually fixed at the point at which the odds ratio is exactly 1:1. In other words, at the point at which there is equal evidence for a word being old and a word being new. Thus changes in the criterion are not cited as being reasons why endorsement rates change. Therefore, the data favour criterion shift explanations of the mirror effect over differentiation explanations.

Intentionality is another area in which the classification and recognition tasks differ. In classification, endorsing NC words is required by the task, yet in recognition endorsing NC words is inappropriate. That NC words are endorsed at all in recognition implies that information about the rule consistency of words is used to aid recognition decisions in addition to explicit recollection and familiarity based on episodic factors. Even if episodic familiarity is simply a weaker form of recollection (Jimenez et al., 1996), the same is not true of structural familiarity. If it were, then there should be more verbalisable knowledge of the rule set with strong-study rather than weak-study conditions, which did not happen. At any rate, intentionality seems irrelevant to the structural effect – regardless of intent, NC words were endorsed more than NI words in both recognition and classification. However, intentionality is clearly having some effect as the decision criterion shifted from recognition to classification. This could reflect participants using their lack of verbalisable knowledge of the rule set as a guide to setting their criterion. Because they could not verbalise the rule set, they judged the task as difficult and set a liberal criterion in classification. At least the setting of a criterion appears to be a controllable action.

The results are also relevant to debates in signal-detection models. The strength-of-evidence axis can be taken to represent either a single process (C. J. Berry, Shanks, & Henson, 2008) or multiple processes combining into one strength-of-evidence dimension (Wixted & Mickes, 2010). The evidence presented here favours the multiple process

account. In a single-process account there would not be a dissociable influence of study strength on the episodic and structural effects. Although C. J. Berry, Shanks and Henson (2008) cite differences in the noise levels affecting each distribution, the analyses in Chapter 5 argue against this as an explanation in this case. Multiple-process accounts of signal detection can simply state that study strength does not affect the process that feeds the structural effect. Comparing recognition and classification results also suggests another area that may be under intentional control of participants – the extent to which each influence is weighted when contributing to the strength of evidence. HRs were lower in classification than in recognition and, as discussed above, this could reflect participants relying less on recollection in classification than in recognition. From a Wixted and Mickes style signal-detection perspective, this directly translates to a lower weighting of the contribution of recollection to the total strength-of-evidence. Further investigation of the notion of intentionality would be of interest in terms of signal-detection models – if the strength of evidence is made up of separate familiarity and recollection distributions and these distributions in turn are made up of several different factors, then which factors can participants intentionally emphasise and deemphasise and which factors are involuntary? The traditional view is that recollection is explicit and controllable and familiarity is not, but is it possible that if there are different types of familiarity contributing to the familiarity distribution that some types are, in fact, controllable?

6.3 Limitations and future research

In this section, further limitations and future research possibilities are discussed. In Chapter 5, bigrams were discounted as a possible basis for the structural effect. However, it is possible that participants learn other features of the stimuli that are correlated with the rule set. One factor which could be cited is word length, but word length is highly correlated with bigram frequency. The analysis that ruled out bigram frequency thus also rules out word length as a factor in the structural effect. What other features participants might be learning is unclear, and so this limitation is not in and of itself strong enough to invalidate the findings from the previous chapters.

Another limitation is the choice of participants. All the experiments here were conducted on primarily English speaking, university students in early adulthood. Other populations might produce different results. For instance, older adults tend to have poorer recollection than younger adults but relatively unimpaired familiarity (Yonelinas, 2002). It would be instructive to see if structural familiarity is also spared in older populations, or if it

is impaired. Education could also play a role, in that a general understanding of the different ways of classifying words might be required to learn a rule set that depends on such classifications. Similarly, different languages might have different ways of classifying words which could interfere with the specific rule set used for the experiments presented here. Examining such population differences could lead to greater insight into the structural effect and would be worth pursuing.

A.S. Reber (1989) demonstrated that people's learning of AG was idiosyncratic by showing the schematic used to create the stimuli to participants at different points during learning. This disrupted any learning that had been achieved up to that point. The conjunctive rule-set used here is not the same as an AG network, but it is still possible that people learn it in an idiosyncratic way, and if this is the case it may be that the awareness questionnaire classified people as unaware of the rule set when in fact they were aware. For instance a participant could have been choosing words that were rare and hard to touch, but not associate the "hard to touch" rule with concreteness. Additionally, they may have forgotten the rule they used by the time they answered the questionnaire. Two experimental changes could address this issue. Participants could be guided to attend to particular features of the stimuli at different stages throughout the study phase to try to produce a similar method of learning across all participants. This might serve as a similar manipulation to A.S. Reber's schematic, depending on which features are emphasised. Additionally a think-aloud requirement could be added in order to elicit the participants' thoughts at the time of their choice. P.J. Reber and Kotovsky (1997) asked participant to think aloud whilst solving a puzzle to demonstrate that the participants did not have verbalisable knowledge of the task, although using the think-aloud task reduced performance on the primary puzzle task.

The fact that no manipulation could increase the structural effect means there are still unanswered questions as to what exactly is learned in implicit learning. There are several possible explanations of the structural effect being insensitive to study strength that were not investigated in the experiments presented here. Runger and Frensch (2008) demonstrated that unexpected events encouraged deeper processing. Hintzman and Curran (1995) reached a similar conclusion. Thus it may be the repeated words at study were an expected event and so only supported shallow processing of word frequency and not the deeper processing required to analyse word features. An experiment in which the additional words presented at study were different words rather than repetitions of the already presented words could help to investigate this explanation. Longer study lists are sometimes associated with lower recognition performance (e.g. Gronlund & Elam, 1994), so it would be instructive to see how

the structural effect responded to such a manipulation. Alternatively the repeated words could be modified in some way as to present them in an unexpected manner, which may also result in their full features being processed. It could also be that study strength does increase the degree of learning of the structural effect, but something about the tasks prevented the expression of that learning. Again, a manipulation that made particular features salient at test could help to clarify this point. Reducing the number of words at study would also be of interest. If a point could be reached at which there were too few words to support a structural effect, it is possible that an increase in repetitions would serve to produce a structural effect. Although this would not be the same as showing that the magnitude of the structural effect is sensitive to study strength, it would clarify whether increasing repetitions has the effect of stabilising the structural effect at a particular level.

A further manipulation could involve intentionality. The results of the experiments suggested that participants cannot control the expression of the structural effect. However, this could easily be tested by informing participants in a recognition task of the existence of the rule set and that past participants have mistaken rule-consistent words for old words. Presumably participants must experience a higher strength of evidence for NC words than for NI words, and it might be that without something to attribute this feeling to they interpret it as being due to oldness. If they can attribute this strength of evidence to rule consistency then despite a lack of verbalisable knowledge, participants may still be able to exert intentional control. In this case, the structural effect may disappear.

More generally, research such as this informs us about how we interact with the world. The structural effect may be insensitive to study strength but that is clearly not the case for all learning. Understanding what types of learning benefits from repetition and what learning does not could inform decisions about how best to teach and test different kinds of knowledge. A deeper understanding of how implicit learning operates would pay similar dividends. The research presented here contributes to that body of knowledge.

6.4 Final Conclusion

Breaking out of established methods such as AG experiments can provide new insights into how we learn complicated rules and apply them in everyday life. Any good model of memory and learning will have to encompass the pattern of data demonstrated regarding the structural effect. In order to do this, we must break down the walls that separate two literatures that, by rights, belong together – those of implicit learning and recognition memory.

This page intentionally left blank

Appendices

Appendix A

Words used in Experiments 1 and 2

CC = Common-concrete CA = common-abstract RC = rare-concrete RA = rare-abstract

ARM	CC	GAS	CC
BALL	CC	HALL	CC
BAR	CC	HOME	CC
BODY	CC	HOUSE	CC
BRIDGE	CC	ISLAND	CC
BUILDING	CC	MACHINE	CC
CHURCH	CC	MATERIAL	CC
EARTH	CC	MONEY	CC
EQUIPMENT	CC	OFFICE	CC
FACE	CC	ROOM	CC
FARM	CC	SECRETARY	CC
FIRE	CC	WATER	CC
HAIR	CC	BATTLE	CC
HOTEL	CC	BILL	CC
HUMAN	CC	BOOK	CC
JACK	CC	APARTMENT	CC
KING	CC	ARMY	CC
LADY	CC	BANK	CC
LAND	CC	BOY	CC
MARKET	CC	CAR	CC
MOTHER	CC	CASE	CC
OFFICER	CC	CATTLE	CC
PAPER	CC	CITY	CC
PLANE	CC	CLAY	CC
POET	CC	CLOTHES	CC
POST	CC	DOOR	CC
PRESIDENT	CC	DRINK	CC
RIVER	CC	EYE	CC
ROAD	CC	FOOD	CC
ROSE	CC	GIRL	CC
SHIP	CC	GLASS	CC
SPRING	CC	GUN	CC
STREET	CC	HEAD	CC
STUDENT	CC	HEART	CC
SUN	CC	HORSE	CC
TEACHER	CC	ABBREVIATION	RA
TRAIN	CC	ALLURE	RA
UNIVERSITY	CC	APPLICATOR	RA
WINDOW	CC	BELIEVABLE	RA
WOMAN	CC	BITTERSWEET	RA
CENT	CC	BOOSTER	RA
CHILDREN	CC	COMMENDATION	RA
DINNER	CC	DEHUMANIZE	RA
FATHER	CC	DEPICTION	RA
FILM	CC	DIABOLICAL	RA

DIVIDER	RA	AGNOMEN	RA
EXHAUSTIBLE	RA	ANIMISM	RA
GLORIFICATION	RA	BESMIRCH	RA
INDESTRUCTIBLE	RA	BROWSING	RA
INKLING	RA	BY-PASS	RA
LARVAL	RA	CAPRICIOUS	RA
MELD	RA	CONDUIT	RA
MEDITATE	RA	CONTORTION	RA
NAB	RA	CORTEGE	RA
OBSCENITY	RA	DAPPLED	RA
PENSIVE	RA	DEEM	RA
PRESCRIPTIVE	RA	DETACH	RA
PUNITIVE	RA	FEATHERY	RA
RECREATE	RA	FLOATER	RA
ROVE	RA	FOREPART	RA
SARCASM	RA	FROSTED	RA
SATIATE	RA	HARPY	RA
SCATHING	RA	HELIOCENTRIC	RA
SHAKESPEARIAN	RA	IMBECILE	RA
SECLUDE	RA	LEATHERY	RA
SULTRY	RA	WOMB	RC
SYNTHESIZE	RA	WIG	RC
SUP	RA	VEAL	RC
THANKLESS	RA	UMPIRE	RC
TIDAL	RA	TORNADO	RC
TREASONOUS	RA	THIMBLE	RC
UNDERRATE	RA	STOCKING	RC
UNFLATTERING	RA	SQUIRREL	RC
WASTAGE	RA	SNAIL	RC
UNHOOK	RA	SHRUB	RC
ARTIFICIALITY	RA	SAP	RC
AVERT	RA	RASPBERRY	RC
BEMOAN	RA	PROJECTOR	RC
CATEGORIZE	RA	POLIO	RC
CLOD	RA	PICKLE	RC
CUSTODIAL	RA	PERCH	RC
EFFLORESCE	RA	PARCEL	RC
EQUINE	RA	OATMEAL	RC
FINIAL	RA	MORPHINE	RC
GRIZZLED	RA	MAMMAL	RC
HOME-BRED	RA	KITE	RC
INFINITIVE	RA	JEWEL	RC
JUDICIOUS	RA	INFIRMARY	RC
NEBULA	RA	HOCKEY	RC
OVERBLOWN	RA	HELMET	RC
PARTAKING	RA	GROCER	RC
PRAYERFUL	RA	FLORA	RC
SHIFTLESS	RA	ENAMEL	RC
STEELY	RA	DUCHESS	RC
UNSAID	RA	DOE	RC

DIAL	RC	WRITTEN	CA
DANDELION	RC	UNITED	CA
CLOVE	RC	TOWARD	CA
CARROT	RC	THOUSAND	CA
CAMEL	RC	SUPPORT	CA
BOAR	RC	STRESS	CA
AVALANCHE	RC	SON	CA
ASPARAGUS	RC	SOMEONE	CA
APRICOT	RC	SAYING	CA
ACCORDION	RC	PRESS	CA
WAND	RC	POSSIBLE	CA
GULLET	RC	POLICY	CA
HORSEHAIR	RC	ORIGINAL	CA
FOWL	RC	MOREOVER	CA
HADDOCK	RC	MET	CA
SLEET	RC	MEDICAL	CA
MUG	RC	MARKED	CA
MOTH	RC	LOT	CA
GIG	RC	JAZZ	CA
CLARINET	RC	ITSELF	CA
ICICLE	RC	INTERNATIONAL	CA
BAGPIPE	RC	INDICATE	CA
BRACELET	RC	HAVING	CA
EARTHWORM	RC	GIVEN	CA
EWE	RC	FORWARD	CA
PENICILLIN	RC	ENERGY	CA
BELFRY	RC	DRIVE	CA
HARE	RC	DEMAND	CA
SKATE	RC	COVERED	CA
MINER	RC	COMPLETE	CA
CEDAR	RC	CLOSED	CA
KNITTING	RC	CHOICE	CA
MORGUE	RC	BRITISH	CA
CAULIFLOWER	RC	BETTER	CA
PUDDLE	RC	ANYTHING	CA
TYPHOON	RC	ANALYSIS	CA
PLUM	RC	ALLOWED	CA
HALTER	RC	ACTIVE	CA
COWHIDE	RC	ACROSS	CA
JADE	RC	ABOVE	CA
RETAILER	RC	DOUBT	CA
WEED	RC	ACTUAL	CA
PODIUM	RC	DISTRIBUTION	CA
SOOT	RC	BECOME	CA
MOSQUITO	RC	LET	CA
WALRUS	RC	FRENCH	CA
TUNIC	RC	PREVIOUS	CA
OFFAL	RC	MOVING	CA
ZIPPER	RC	SERVE	CA
DRESSER	RC	BELIEVE	CA

CHARGE	CA	AVAILABLE	CA
APPLIED	CA	PROPER	CA
DESIGNED	CA	SIMPLY	CA
THINKING	CA	OFFER	CA
FEED	CA	CANNOT	CA
CIVIL	CA	BASIC	CA
SPENT	CA	PARTICULAR	CA
NATURAL	CA	POPULAR	CA
LOOKING	CA	STATED	CA
EVERYONE	CA	REGARD	CA
DOES	CA	RAISED	CA
ADD	CA	INTERESTED	CA
COMMON	CA	GROWING	CA
OPPOSITE	CA	DATA	CA
HIT	CA		
EQUAL	CA		

Appendix B

Questionnaire used in all experiments. Question 1 varied depending on the task - recognition or classification. This questionnaire was electronically delivered so the tables are formatted to simulate the presentation of the answer boxes. Questions 1, 2 and 3 were answered via free text boxes whilst Questions 4 and 5 required participants to select from a list of alternatives.

Question 1a (recognition condition):

When you were making your decisions about which words you had seen before and which you had not, what criteria other than pure memory did you use to make your decisions?

Question 1a (classification condition):

Did you use any rules to judge the words as either consistent or inconsistent with the rule set as you went through the experiment, and if so, what were they?

Question 2:

The list of words you read in the study phase were included on the list according to a specific set of rules. What do you think these rules were?

Question 3:

Were there any rules that you considered but rejected?

Question 4:

Following is a list of different possible factors that you may have used to help you identify words in the experimental phase. For each one, please indicate by clicking the appropriate button if you did or didn't use it. If you did then please write how exactly you used that factor in the box provided.

Factor	Yes	No	Specific details of use
Length (number of letters)			
Number of syllables			
Category of word (noun/verb/adverb etc)			
Letters			
Meaning			
Familiarity to you			
Frequency (I.e. rare or common)			
Association to other words			
Likely position in a sentence			
Concreteness			

Question 5:

For a word to be included on the study list it had to conform to two separate criteria. Knowing this, which two of the possible factors below do you believe were used as criteria? Select yes for two criteria below and then use the box at the bottom to describe how you think the criteria were related

Criteria	Yes	No
Length (number of letters)		
Number of syllables		
Category of word (noun/verb/adverb etc)		
Letters		
Meaning		
Familiarity to you		
Frequency (I.e. rare or common)		
Association to other words		
Likely position in a sentence		
Concreteness		
How were they related?		

Appendix C

Attributions Analysis from Chapter 4

Attributions Analysis from Experiment 6.

Attributions of familiarity after a participant said old were labelled as episodic familiarity, whilst attributions of consistent were labelled as structural familiarity. Recollection, episodic familiarity and structural familiarity attributions were separately entered into ANOVAs with within-subject factors of word type (old versus NC versus NI) and deadline (short versus long) and between-subject factors of task type (recognition versus classification) and study strength (weak study versus strong study). Responses labelled as guesses were left in to maintain acceptable cell counts.

Recollection.

See Table C.1 for recollection means and standard errors and Table C.2 for the ANOVA results and related pairwise comparisons.

Table C.1

Recollection Endorsements by Word Type, Task Type, Study Strength and Deadline for Experiment 6 (SE in brackets)

Task and word type	Short deadline		Long deadline	
	Weak study	Strong study	Weak study	Strong study
Recognition				
Old	.36 (.04)	.57 (.07)	.37 (.03)	.61 (.05)
NC	.15 (.05)	.13 (.06)	.14 (.03)	.06 (.03)
NI	.10 (.03)	.05 (.02)	.11 (.03)	.04 (.02)
Classification				
Old	.29 (.05)	.44 (.05)	.36 (.04)	.48 (.05)
NC	.03 (.01)	.04 (.01)	.04 (.01)	.01 (.00)
NI	.04 (.02)	.00 (.00)	.04 (.02)	.01 (.01)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table C.2

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Deadline) ANOVA on Recollection Attributions from Experiment 6

Effect	*F* value (df)	*p* value	η^2
Word type	$F(2, 104) = 319.37$	< .001*	.86
Old versus NC words	$F(1, 52) = 23.58$	< .001*	.30
NC versus NI words	$F(1, 52) = 25.14$	< .001*	.31
Task type	$F(1, 52) = 8.74$	.005*	.14
Word type by study strength	$F(2, 104) = 27.24$	< .001*	.34
Study strength for old words	$F(1, 52) = 15.44$	< .001*	.23
Study strength for NC words	$F(1, 52) = 1.54$	.22	-
Study strength for NI words	$F(1, 52) = 5.25$	.03*	.09
Word type and deadline	$F(2,104) = 7.43$	< .001*	.12
Deadline for old words	$F(1, 52) = 5.62$	.02*	.10
Deadline for NC and NI words	highest $F(1, 52) = 3.49$	.07	-
There were no other significant effects	$F(2, 104) = 2.14$	.12	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction. * = *p* value denotes significant difference.

There were more recollection attributions for old (*M* = .43, *SE* = .02) than for NC words (*M* = .07, *SE* = .01) and more for NC than for NI words (*M* = .05, *SE* = .01). There were more recollection attributions in recognition (*M* = .22, *SE* = .02) than in classification (*M* = .15, *SE* = .02).

Old recollection attributions increased from weak- (*M* = .34, *SE* = .03) to strong-study (*M* = .52, *SE* = .03) conditions, NC recollection attributions did not change from weak- (*M* = .09, *SE* = .02) to strong-study (*M* = .06, *SE* = .02) conditions and NI recollection attributions decreased from weak- (*M* = .07, *SE* = .01) to strong-study (*M* = .03, *SE* = .01) conditions.

Only the old word recollection attributions decreased from long-deadline (*M* = .45, *SE* = .02) to short-deadline (*M* = .41, *SE* = .03) conditions. NC recollection attributions did not change from long-deadline (*M* = .06, *SE* = .01) to short-deadline (*M* = .09 *SE* = .02) conditions and NI recollection attributions also did not change from long-deadline (*M* = .05, *SE* = .01) to short-deadline (*M* = .05, *SE* = .01) conditions.

Episodic familiarity.

Episodic familiarity (EF) attributions were analysed in the same way as recollect attributions - see Table C.3 for means and standard errors and Table C.4 for the results of the ANOVA.

Table C.3

Episodic Familiarity Attributions by Word Type, Study Strength, Task Type and Deadline for Experiment 6 (SE in brackets)

Task and word type	Short deadline		Long deadline	
	Weak study	Strong study	Weak study	Strong study
Recognition				
Old	.27(.03)	.18 (.04)	.22 (.03)	.15 (.03)
NC	.36 (.03)	.18 (.04)	.23 (.03)	.09 (.02)
NI	.34 (.04)	.17 (.04)	.19 (.02)	.08 (.01)
Classification				
Old	.16 (.03)	.14 (.02)	.14 (.03)	.15 (.02)
NC	.16 (.03)	.09 (.02)	.12 (.02)	.08 (.02)
NI	.13 (.04)	.03 (.01)	.10 (.02)	.05 (.01)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table C.4

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Deadline) ANOVA on Episodic Familiarity Attributions from Experiment 6

Effect	*F* value (df)	*p* value	η^2
Word type	$F(2, 104) = 6.00$	< .01*	.10
Old versus NI	$F(1, 55) = 7.12$	.01*	.11
NC versus NI	$F(1, 55) = 12.21$	.001*	.18
Old versus NC	$F < 1$		
Deadline	$F(1, 52) = 23.29$	< .001*	.31
Study strength	$F(1, 52) = 18.47$	< .001*	.26
Task type	$F(1, 52) = 21.24$	< .001*	.29
Word type by study strength	$F(2, 104) = 4.23$	.02*	.07
Study strength for old words	$F(1, 52) = 2.32$	.13	-
Study strength for NC words	$F(1, 52) = 18.32$	< .001*	.26
Study strength for NI words	$F(1, 52) = 25.60$	< .001*	.33
Word type by task type	$F(2, 104) = 3.37$	.04*	.06
Task type for old words	$F(1, 52) = 3.84$	.06	-
Task type for NC words	$F(1, 52) = 16.91$	< .001*	.24
Task type for NI words	$F(1, 52) = 30.85$	< .001*	.37
Deadline by task type	$F(1, 52) = 13.92$	< .001*	.21
Deadline for recognition	$F(1, 52) = 35.16$	< .001*	.40
Deadline for classification	$F < 1$	-	-
Study strength by task type	$F(1, 52) = 4.24$	.04*	.07
Study strength for recognition	$F(1, 52) = 19.40$	< .001*	.27
Study strength for classification	$F(1, 52) = 2.61$	.11	-
There were no other significant effects	Highest $F(2, 104) = 3.08$	.052	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.
* = *p* value denotes significant difference.

There were fewer EF attributions for NI words ($M = .13$, $SE = .01$) than for old ($M = .18$, $SE = .01$) or NC words ($M = .17$, $SE = .01$) with old and NC EF attributions being the same. There were more EF attributions with a short-deadline ($M = .19$, $SE = .01$) than with a long-deadline ($M = .13$, $SE = .01$). There were fewer EF attributions in weak-study ($M = .20$, $SE = .01$) than strong-study ($M = .12$ $SE = .01$) conditions. Finally there were

more EF attributions in recognition (M = .20, SE = .01) than in classification (M = .11, SE = .01).

The word-type by study-strength interaction reflected the fact that EF attributions to old words did not change from weak-study (M = .20, SE = .02) to strong-study conditions (M = .16, SE = .02), whilst NC EF attributions reduced from weak-study (M = .22, SE = .02) to strong-study conditions (M = .11, SE = .02). NI EF attributions also reduced from weak-study (M = .19, SE = .01) to strong-study conditions (M = .08, SE = .02).

The word-type by task-type interaction reflected the fact that EF attributions to old words did not change from recognition (M = .20, SE = .02) to classification (M = .15, SE = .02), whilst NC EF attributions decreased from recognition (M = .22, SE = .02) to classification (M = .12, SE = .02). NI EF attributions also decreased from recognition (M = .19, SE = .01) to classification (M = .08, SE = .01).

The deadline by task-type interaction reflected the fact that in recognition there were more EF attributions in short-deadline (M = .25, SE = .02) conditions than in long-deadline (M = .16, SE = .01) conditions, whilst in classification there was no change from short deadline (M = .12, SE = .02) to long deadline (M = .11, SE = .01).

Finally, the study-strength by task-type interaction reflected fewer EF attributions in the weak-study (M = .27, SE = .02) condition than the strong-study (M = .14, SE = .02) condition in recognition, but no change from weak-study (M = .14, SE = .02) to strong-study (M = .09, SE = .02) conditions in classification.

Structural familiarity.

Structural familiarity (SF) attributions were analysed in the same way as recollect attributions. See Table C.5 for means and standard errors and Table C.6 for the results of the ANOVA.

Table C.5

Structural Familiarity Attributions by Word Type, Study Strength, Task Type and Deadline for Experiment 6 (SE in brackets)

Task and word type	Short deadline			Long deadline		
	Weak study	Strong study	Total	Weak study	Strong study	Total
Recognition						
Old	.19 (.04)	.16 (.03)	.17 (.03)	.17 (.02)	.13 (.02)	.15 (.02)
NC	.19 (.04)	.23 (.05)	.21 (.03)	.25 (.04)	.36 (.06)	.30 (.04)
NI	.24 (.04)	.20 (.05)	.22 (.03)	.23 (.03)	.25 (.05)	.25 (.03)
Classification						
Old	.16 (.03)	.09 (.02)	.12 (.02)	.08 (.02)	.12 (.03)	.10 (.02)
NC	.22 (.02)	.28 (.06)	.25 (.03)	.10 (.02)	.24 (.07)	.17 (.03)
NI	.25 (.04)	.27 (.06)	.26 (.03)	.09 (.03)	.21 (.07)	.15 (.03)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table C.6

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Deadline) ANOVA on Structural Familiarity Attributions from Experiment 6

Effect	*F* value (df)	*p* value	η^2
Word type	$F(2, 104) = 21.22$	< .001*	.29
Old versus NI	$F(1, 52) = 20.78$	< .001*	.27
NC versus NI	$F < 1$	-	-
Old versus NC	$F(1, 52) = 24.09$	< .001*	.30
Word type by study strength	$F(2, 104) = 5.93$	.004*	.10
Old versus NI words in strong study	$F(1, 52) = 13.65$	< .001*	.35
NI versus NC words in strong study	$F(1, 52) = 5.84$	< .02*	.19
All word type comparisons in weak study	$F < 1$	-	
Deadline by study strength	$F(1, 52) = 7.82$	.007*	.13
Deadline for weak study	$F(1, 52) = 10.50$	.002*	.17
Deadline for strong study	$F < 1$	-	-
Deadline by task type	$F(1, 52) = 16.67$	< .001*	.24
Deadline for recognition	$F(1, 52) = 2.93$	.09	-
Deadline for classification	$F(1, 52) = 16.91$	< .001*	.24
Word type by deadline	$F(2, 104) = 3.46$	.03*	.06
Old versus NC for short deadline	$F(1, 55) = 13.47$	< .001*	.20
Old versus NI for short deadline	$F(1, 55) = 117.07$	< .001*	.24
NC versus NI for short deadline	$F < 1$	-	-
Old versus NI for long deadline	$F(1, 55) = 14.73$	< .001*	.21
NI versus NC for long deadline	$F(1, 55) = 6.07$	.02*	.10
Word type by deadline by task type	$F(2, 104) = 10.80$	< .001*	.17
Deadline for recognition old attributions	$F < 1$	-	-
Deadline for recognition NC attributions	$F(1, 52) = 13.99$	< .001*	.21
Deadline for recognition NI attributions	$F(1, 52) = 1.01$	.32	-
Deadline for classification old attributions	$F(1, 52) = 1.62$	.21	-
Deadline for classification NC attributions	$F(1, 52) = 11.43$	.001*	.18
Deadline for classification NI attributions	$F(1, 52) = 19.95$	< .001*	.28
There were no other significant effects	$F(1, 52) = 2.04$	.16	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.

* = *p* value denotes significant difference.

There were fewer SF attributions for old words (M = .13, SE = .01) than for NC (M = .23, SE = .02), and NI words (M = .22, SE = .02).

The word-type and study-strength interaction was due to there being no difference between old (M = .15, SE = .02), NC (M = .19, SE = .03) and NI (M = .21, SE = .03) SF attributions in the weak-study condition, whilst in the strong-study condition there were fewer old SF attributions (M = .12, SE = .02) than there were NI SF attributions (M = .23, SE = .03), and there were fewer NI SF attributions than there were NC SF attributions (M = .28, SE = .03). The interaction did not reflect any general increases in attributions by study strength.

The deadline and study-strength interaction reflected more SF attributions with short deadline (M = .21, SE = .02) than with long deadline (M = .15, SE = .02) in the weak-study condition, whilst in the strong-study condition there was no difference from short-deadline (M = .20, SE = .03) to long-deadline conditions (M = .22, SE = .03).

The deadline by task-type interaction reflected a reduction in SF attributions from short-deadline (M = .21, SE = .03) to long-deadline conditions (M = .14, SE = .02) in the classification task, whilst there was no change from short-deadline (M = .20 SE = .03) to long-deadline (M = .23, SE = .03) conditions in recognition.

The word-type by deadline interaction reflected fewer old SF attributions (M = .15, SE = .02) than NC (M = .23, SE = .02), or NI SF attributions (M = .24, SE = .02), in the short deadline whilst at the long deadline there were fewer old SF attributions (M = .12, SE = .01) than there were NI SF attributions (M = .20, SE = .02), and there were fewer NI SF attributions than there were NC SF attributions (M = .24, SE = .02).

The word-type, deadline and task-type interaction was due to the fact that in recognition deadline does not change old or NI SF attributions, but NC SF attributions increased from short-deadline to long-deadline conditions. In classification, again old SF attributions did not change with deadline, whilst NC and NI SF attributions both decreased from short- to long-deadline conditions.

Attributions Analysis from Experiment 7

Recollection.

Recollect attributions were analysed in the same way as the recollect attributions from Experiment 6 except instead of a within-subject factor of deadline there was a between-subject factor of distraction (distracted versus not distracted). For means and standard errors see Table C.7 and for the results of the ANOVA see Table C.8.

Table C.7

Recollection Endorsement Rates by Task Type, Word Type, Study Strength and Distraction from Experiment 7 (SE in brackets)

Task and word type	Not Distracted		Distract		Total	
	Weak study	Strong study	Weak study	Strong study	Weak study	Strong study
Recognition						
Old	.40 (.04)	.65 (.04)	.47 (.03)	.64 (.05)	.44 (.03)	.65 (.03)
NC	.09 (.03)	.03 (.01)	.18 (.03)	.06 (.01)	.14 (.01)	.04 (.01)
NI	.06 (.02)	.02 (.01)	.11 (.02)	.05 (.01)	.08 (.01)	.04 (.01)
Classification						
Old	.39 (.03)	.55 (.05)	.40 (.05)	.51 (.05)	.39 (.03)	.53 (.03)
NC	.08 (.02)	.05 (.01)	.08 (.03)	.05 (.01)	.08 (.01)	.05 (.01)
NI	.05 (.01)	.02 (.01)	.05 (.01)	.03 (.01)	.05 (.01)	.03 (.01)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table C.8

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Distraction) ANOVA on Recollection Attributions from Experiment 7

Effect	*F* value (df)	*p* value	η^2
Word type	$F(2, 238) = 780.47$	< .001*	.89
Old versus NC	$F(1, 126) = 503.10$	< .001*	.80
NC versus NI	$F(1, 126) = 41.91$	< .001*	.25
Task type	$F(1, 119) = 8.63$	.004*	.88
Word type by study strength	$F(2, 238) = 50.97$	< .001*	.30
Study strength for old words	$F(1, 119) = 31.30$	< .001*	.21
Study strength for NC words	$F(1, 119) = 18.41$	< .001*	.13
Study strength for NI words	$F(1, 119) = 11.63$	.001*	09
Word type by task type by study strength	$F(2, 238) = 3.84$	.02*	.03
Study strength for recognition old words	$F(1, 119) = 24.52$	< .001*	.17
Study strength for recognition NC words	$F(1, 119) = 20.30$	< .001*	.15
Study strength for recognition NI words	$F(1, 119) = 11.09$	.001*	.08
Study strength for classification old words	$F(1, 119) = 9.01$	.003*	.07
Study strength for classification NC words	$F(1, 119) = 2.63$	.11	-
Study strength for classification NI words	$F(1, 119) = 2.34$	.13	-
There were no other significant effects	$F(2, 238) = 3.31$	.07	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction. * = *p* value denotes significant difference.

There were more old recollection endorsements (M = .50, SE = .01) than NC recollection endorsements (M = .08, SE = .01) and more NC endorsements than NI recollection endorsements (M = .05, SE = .01). There were more recollection endorsements in recognition (M = .23, SE = .01) than in classification (M = .19, SE = .01).

The word-type and study-strength interaction reflected the fact that old recollection endorsements increased from weak-study (M = .41, SE = .02) to strong-study conditions (M = .58, SE = .02), whilst NC decreased from weak-study (M = .11, SE = .01) to strong-study conditions (M = .04, SE =.01). NI also decreased from weak-study (M = .07, SE = .01) to strong-study conditions (M = .03, SE = .01).

The word-type, task-type and study-strength interaction reflected an increase in old recollection endorsements and a decrease in NC and NI recollection endorsements from weak- to strong-study conditions in recognition. However, in classification, old recollection endorsements increased, but NC and NI recollection endorsements did not change from weak-study to strong-study conditions.

Episodic familiarity.

EF endorsements were analysed in the same way as recollection endorsements. For means and standard errors Table C.9 and for the results of the ANOVA see Table C.10.

Table C.9

Episodic Familiarity Attributions by Word Type, Study Strength, Task Type and Distraction for Experiment 7 (SE in brackets)

Task and word type	Not Distracted			Distracted		
	Weak study	Strong study	Total	Weak study	Strong study	Total
Recognition						
Old	.34 (.02)	.21 (.03)	.27 (.02)	.24 (.02)	.20 (.03)	.22 (.02)
NC	.30 (.02)	.14 (.03)	.22 (.01)	.23 (.02)	.14 (.02)	.19 (.01)
NI	.24 (.02)	.10 (.03)	.17 (.02)	.22 (.02)	.11 (.02)	.16 (.02)
Classification						
Old	.22 (.03)	.14 (.02)	.18 (.02)	.20 (.03)	.16 (.02)	.18 (.02)
NC	.16 (.02)	.08 (.02)	.12 (.01)	.14 (.02)	.08 (.02)	.11 (.02)
NI	.16 (.03)	.07 (.02)	.12 (.02)	.10 (.01)	.05 (.02)	.07 (.02)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table C.10

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Distraction) ANOVA on Episodic Familiarity Attributions from Experiment 7

Effect	*F* value (df)	*p* value	η^2
Word type	$F(2, 238) = 50.26$	< .001*	.30
Old versus NC	$F(1, 126) = 41.91$	< .001*	.25
NC versus NI	$F(1, 126) = 23.29$	< .001*	.16
Task type	$F(1, 119) = 30.74$	< .001*	.20
Study strength	$F(1, 119) = 44.01$	< .001*	.27
Word type by task type by distraction	$F(2, 238) = 3.90$	.02*	.03
Distraction for recognition old words	$F(1, 119) = 4.41$	.04 *	.04
All other comparisons	$F(1, 119) = 3.26$	.07	-
There were no other significant effects	$F(1, 119) = 3.30$	.07	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction.
* = *p* value denotes significant difference.

There were more old EF endorsements ($M = .21$, $SE = .01$) than NC EF endorsements ($M = .16$, $SE = .01$), and more NC EF endorsements than NI EF endorsements ($M = .13$, $SE = .01$). There were more EF endorsements in recognition (*M*

= .21, *SE* = 01) than in classification (*M* = .13, *SE* = .01). There were fewer EF endorsements in the strong-study (*M* = .12, *SE* = .01) than in the weak-study condition (*M* = .21, *SE* = .01).

The word-type, task-type and distraction interaction reflected a decrease in old EF endorsements from not-distracted to distracted in recognition conditions, but only in recognition. Distraction had no other effect in recognition or classification.

Structural familiarity.

SF endorsements were analysed in the same way as recollection endorsements. See Table C.11 for means and standard errors and Table C.12 for the results of the ANOVA.

Table C.11

Structural Familiarity Attributions by Word Type, Study Strength, Task Type and Distraction for Experiment 7 (SE in brackets)

Task and word type	Not Distracted		Distracted		Total		
	Weak study	Strong study	Weak study	Strong study	Weak study	Strong study	Total
Recognition							
Old	.14 (.02)	.08 (.01)	.09 (.01)	.05 (.01)	.12 (.01)	.07 (.01)	.09 (.01)
NC	.27 (.04)	.35 (.04)	.19 (.02)	.25 (.04)	.23 (.02)	.30 (.02)	.27 (.02)
NI	.24 (.03)	.30 (.03)	.22 (.03)	.20 (.03)	.23 (.02)	.25 (.02)	.24 (.02)
Classification							
Old	.15 (.02)	.10 (.02)	.15 (.02)	.14 (.03)	.15 (.02)	.12 (.01)	.14 (.01)
NC	.25 (.03)	.24 (.02)	.21 (.03)	.19 (.03)	.23 (.02)	.22 (.02)	.22 (.02)
NI	.23 (.03)	.17 (.03)	20 (.03)	.15 (.03)	.21 (.02)	.16 (.02)	.19 (.02)

Note. Old = words seen at study, NC = New rule-consistent words, NI = New rule-inconsistent words.

Table C.12

Results of 3 (Word Type) x 2 (Study Strength) x 2 (Task Type) x 2 (Distraction) ANOVA on Structural Familiarity Attributions from Experiment 7

Effect	*F* value (df)	*p* value	η^2
Word type	$F(2, 238) = 92.61$	< .001*	.44
Old versus NI	$F(1, 126) = 114.60$	< .001*	.48
NI versus NC	$F(1, 126) = 17.03$	< .001*	.12
Distraction	$F(1, 119) = 5.14$,	.02*	.04
Word type by task type	$F(2, 238) = 13.56$	< .001*	.10
Task type for old words	$F(1, 119) = 8.65$	.004*	.07
Task type for NC words	$F(1, 119) = 3.07$	.08	-
Task type for NI words	$F(1, 119) = 5.17$	.02*	.04
Word type by study strength	$F(2, 238) = 5.63$	.004*	.04
Study strength for old words	$F(1, 119) = 6.80$	.01*	.05
Study strength for NC and NI words	highest $F(1, 119) = 1.38$	.24	-
Word type by distraction	$F(2, 238) = 4.05$	.02*	.03
Distraction for NC words	$F(1, 119) = 8.12$	.005*	.06
Distraction for Old and NI words	highest $F(1, 119) = 3.25$	.07	-
Word type by study strength by task type	$F(2, 238) = 3.78$	.02*	.03
Study strength for recognition old words	$F(1, 119) = 5.70$	.02*	.05
Study strength for recognition NC words	$F(1, 119) = 4.47$	.04*	.04
Study strength for recognition NI words and classification old, NC and NI words	highest $F(1, 119) = 2.26$	-	-
There were no other significant effects	highest $F(1, 119) = 1.82$	.18	-

Note. Only significant effects are reported. Pairwise comparisons indented below the relevant interaction. * = *p* value denotes significant difference.

There were more SF endorsements for NC words ($M = .24$, $SE = .01$) than for NI words ($M = .21$, $SE = .01$), and more SF endorsements for NI words than for old words ($M = .11$, $SE = .01$). There were more SF endorsements when not distracted ($M = .21$, $SE = .01$) than when distracted ($M = .17$, $SE = .01$).

The word-type by task-type interaction indicated that SF endorsements to old words increased from recognition to classification, SF endorsements to NI words decreased from recognition to classification, whilst NC SF endorsements did not change from recognition to classification.

The word-type by study-strength interaction reflected the fact that SF endorsements to old words decreased from weak-study ($M = .13$, $SE = .01$) to strong-

study conditions (*M* = .09, *SE* = .01), while NC SF endorsements do not change from weak-study (*M* = .23, *SE* = .02) to strong-study conditions (*M* = .26, *SE* = .02) and NI SF endorsements also did not change from weak-study (*M* = .22, *SE* = .02) to strong-study conditions (*M* = .21, *SE* = .02).

The distraction by word-type interaction reflected the fact that NC SF endorsements decreased from not distracted (*M* = .28, *SE* = .02) to distracted (*M* = .21, *SE* = .02), whilst old SF endorsements stay the same from not distracted (*M* = .12, *SE* = 01) to distracted (*M* = .11, *SE* = .01) and NI SF endorsements also did not change from not distracted (*M* = .23, *SE* = .02) to distracted conditions (*M* = .19, *SE* = .02).

The study-strength by word-type by task-type interaction reflected the fact that in recognition, old SF endorsements increased from weak-study to strong-study conditions, and NC SF endorsements decreased from weak-study to strong-study conditions. NI SF endorsements and old, NC endorsements in classification did not change by study strength.

This page intentionally left blank

References

Abrahamse, E. L., van der Lubbe, R. H. J., & Verwey, W. B. (2008). Asymmetrical learning between a tactile and visual serial RT task. *The Quarterly Journal of Experimental Psychology, 61*(2), 210-217.

Abrahamse, E. L., & Verwey, W. B. (2008). Context dependent learning in the serial RT task. *Psychological Research/Psychologische Forschung, 72*(4), 397-404.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38*(4), 341-380.

Antony, S., & Santhanam, R. (2007). Could the use of a knowledge-based system lead to implicit learning? *Decision Support Systems, 43*(1), 141-151.

Arndt, J., & Reder, L. M. (2002). Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(5), 830-842.

Azzopardi, P., & Cowey, A. (1997). Is blindsight like normal near-threshold vision. *Proceedings of the National Academy of Sciences of the United States of America, 94*(25), 14190-14194.

Baddeley, A., Lewis, V., Eldridge, M., & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General, 113*(4), 518-540.

Barnhardt, T. M., & Geraci, L. (2008). Are awareness questionnaires valid? Investigating the use of posttest questionnaires for assessing awareness in implicit memory tests. *Memory & Cognition, 36*(1), 53-64.

Bennett, I. J., Howard, J. H., & Howard, D. V. (2007). Age-related differences in implicit learning of subtle third-order sequential structure. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences, 62*(2), P98-P103.

Berry, C. J., Shanks, D. R., & Henson, R. N. A. (2008). A unitary signal-detection model of implicit and explicit memory. *Trends in Cognitive Sciences, 12*(10), 367-373. doi: 10.1016/j.tics.2008.06.005

Berry, D. C. (1997). *How implicit is implicit learning?* New York, NY US: Oxford University Press.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79*, 251-272.

Bright, J. E. H., & Burton, A. M. (1994). Past midnight: semantic processing in an implicit learning task. *The Quarterly of Experimental Psychology, 47A*(1), 71-89.

Brody, N. (1989). Unconscious learning of rules - comment. *Journal of Experimental Psychology-General, 118*(3), 236-238.

Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: General, 120*(3), 316-323.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology, 29*, 461-473.

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(4), 587-599.

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition, 37*(6), 807-818.

Buchner, A. (1994). Indirect effects of synthetic grammar learning in an identification task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 550-566.

Buchner, A., & Wippich, W. (1998). Differences and commonalities between implicit learning and implicit memory. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning.* (pp. 3-46). Thousand Oaks, CA US: Sage Publications, Inc.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*(2), 231-248.

Channon, S., Shanks, D., Johnstone, T., Vakili, K., Chin, J., & Sinclair, E. (2002). Is implicit learning spared in amnesia? Rule abstraction and item familiarity in artificial grammar learning. *Neuropsychologia, 40*(12), 2185-2197.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception and Psychophysics, 36*(4), 387-395.

Cheesman, J., & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology, 40*(4), 343-367.

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *The Cambridge handbook of computational psychology.* (pp. 396-421). New York, NY US: Cambridge University Press.

Coltheart, M. (1981). MRC Psycholinguistic database user manual: version 1.

Craik, F. I. (1982). Selective changes in encoding as a function of reduced processing capacity. In F. Klix, J. Hoffman & E. Van der Meer (Eds.), *Cognitive research in psychology* (pp. 152-161). Berliong: Deutscher Verlag der Wissenschaffen.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55*(4), 461-478.

DeCarlo, L. T. (2003). Using the PLUM procedure of SPSS to fit unequal variance and generalized signal detection models. *Behavior Research Methods, Instruments & Computers, 35*(1), 49-56.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452-478.

Didierjean, A. (2007). Parity effects in an implicit-learning task. *British Journal of Psychology, 98*, 529-545.

Dienes, Z. (2004). Assumptions of subjective measures of unconscious mental states: higher order thoughts and bias. *Journal of Consciousness Studies, 11*, 25-45.

Dienes, Z. (2008a). Subjective measures of unconscious knowledge. *Progress in Brain Research*(168), 49-64.

Dienes, Z. (2008b). *Understanding psychology as a science: An introduction to scientific and statistical inference*: Palgrave Macmillan.

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1322-1338.

Dienes, Z., & Berry, D. C. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review, 4*(1), 3-23.

Dienes, Z., Broadbent, D., & Berry, D. C. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 875-887.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22*, 735-808.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research/Psychologische Forschung, 69*(5-6), 338-351. doi: 10.1007/s00426-004-0208-3

Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition: An International Journal, 19*(2), 674-681. doi: 10.1016/j.concog.2009.09.009

Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General, 113*(4), 541-555.

Evans, S., & Azzopardi, P. (2007). Evaluation of a 'bias-free' measure of awareness. *Spatial Vision, 20*(1-2), 61-77.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Feeney, J. J., Howard, J. H., & Howard, D. V. (2002). Implicit learning of higher order sequences in middle age. *Psychology and Aging, 17*(2), 351-355.

Foerde, K., Poldrack, R. A., & Knowlton, B. J. (2007). Secondary-task effects on classification learning. *Memory & Cognition, 35*(5), 864-874.

Frensch, P. A. (1998). One concept, multiple meanings: On how to define the concept of implicit learning. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning.* (pp. 47-104). Thousand Oaks, CA US: Sage Publications, Inc.

Frensch, P. A., Buchner, A., & Lin, J. (1994). Implicit learning of unique and ambiguous serial transitions in the presence and absence of a distractor task. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(3), 567-584.

Frensch, P. A., Lin, J., & Buchner, A. (1998). Learning versus behavioral expression of the learned: The effects of a secondary tone-counting task on implicit learning in the serial reaction task. *Psychological Research-Psychologische Forschung, 61*(2), 83-98.

Fu, Q. F., Fu, X. L., & Dienes, Z. (2008). Implicit sequence learning and conscious awareness. *Consciousness and Cognition, 17*(1), 185-202.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type-2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10*(4), 843-876.

Gardiner, J. M., Kaminska, Z., Dixon, M., & Java, R. I. (1996). Repetition of previously novel melodies sometimes increases both remember and know responses in recognition memory. *Psychonomic Bulletin & Review, 3*(3), 366-371.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*(1), 8-20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 5-16.

Glanzer, M., Adams, J. K., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 81-93.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 2*(1), 21-31.

Glanzer, M., Kim, K., & Adams, J. K. (1998). Response distribution as an explanation of the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 633-644.

Goldman, B. L., Martin, E. D., Calamari, J. E., Woodard, J. L., Chik, H. M., Messina, M. G., . . . Wiegartz, P. S. (2008). Implicit learning, thought-focused attention and obsessive-compulsive disorder: A replication and extension. *Behaviour Research and Therapy, 46*(1), 48-61.

Gomez, R. L., Gerken, L., & Schvaneveldt, R. W. (2000). The basis of transfer in artificial grammar learning. *Memory & Cognition, 28*(2), 253-263.

Goschke, T., & Bolte, A. (2007). Implicit learning of semantic category sequences: Response-independent acquisition of abstract sequential regularities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 394-406.

Greene, R. L., & Thapar, A. (1994). Mirror effect in frequency discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 946-952.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1355-1369. doi: 10.1037/0278-7393.20.6.1355

Gruppuso, V., Lindsay, D. S., & Kelley, C. M. (1997). The process-dissociation procedure and similarity: Defining and estimating recollection and familiarity in

recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(2), 259-278. doi: 10.1037/0278-7393.23.2.259

Heuer, H., & Klein, W. (2003). One night of total sleep deprivation impairs implicit learning in the serial reaction task, but not the behavioral expression of knowledge. *Neuropsychology, 17*(3), 507-516.

Hicks, J. L., & Marsh, R. L. (2000). Toward specifying the attentional demands of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1483-1498. doi: 10.1037/0278-7393.26.6.1483

Higham, P. A. (1997a). Chunks are not enough: The insufficiency of feature frequency-based explanations of artificial grammar learning. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Experimentale, 51*(2), 126-138.

Higham, P. A. (1997b). Dissociations of grammaticality and specific similarity effects in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 1029-1045.

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30*(1), 67-80.

Higham, P. A. (2007). No Special K! A signal-detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General, 136*(1), 1-22.

Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. *The Quarterly of Experimental Psychology, 50A*(1).

Higham, P. A., Bruno, D., & Perfect, T. J. (2010). Effects of study list composition on the word frequency effect and metacognitive attributions in recognition memory. *Memory, 18*(8), 883-899.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 57-80.

Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory and Language, 52*, 595-617.

Higham, P. A., & Vokey, J. R. (1994). Recourse to stored exemplars is not necessarily explicit: A comment on Knowlton, Ramus, and Squire (1992). *Psychological Science, 5*(1), 59-60.

Higham, P. A., & Vokey, J. R. (2004). Illusory recollection and dual-process models of recognition memory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 57*(4), 714-744.

Higham, P. A., Vokey, J. R., & Pritchard, J. L. (2000). Beyond dissociation logic: Evidence for controlled and automatic influences in artificial grammar learning. *Journal of Experimental Psychology-General, 129*(4), 457-470.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers, 16*(2), 96-101.

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

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

Hintzman, D. L., & Curran, T. (1995). When encoding fails: Instructions, feedback, and registration without learning. *Memory & Cognition, 23*(2), 213-226.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 302-313.

Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition, 35*(4), 679-688.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513-541.

Jacoby, L. L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(1), 3-22.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*(3), 306-340.

Jacoby, L. L., Jones, T. C., & Dolan, P. O. (1998). Two effects of repetition: Support for a dual-process model of knowledge judgments and exclusion errors. *Psychonomic Bulletin & Review, 5*(4), 705-709.

Jamieson, R. K., Holmes, S., & Mewhort, D. J. K. (2010). Global similarity predicts dissociation of classification and recognition: Evidence questioning the implicit–explicit learning distinction in amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1529-1535. doi: 10.1037/a0020598

Jamieson, R. K., & Mewhort, D. J. K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: An analysis using information theory. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(1), 9-23.

Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *The Quarterly Journal of Experimental Psychology, 62*(3), 550-575.

Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction-time task: Anticipating from experience. *The Quarterly Journal of Experimental Psychology, 62*(9), 1757-1783.

Jiang, Y., & Song, J. H. (2005). Hyperspecificity in visual implicit learning: Learning of spatial layout is contingent on item identity. *Journal of Experimental Psychology-Human Perception and Performance, 31*(6), 1439-1448.

Jimenez, L. (2008). Taking patterns for chunks: Is there any evidence of chunk learning in continuous serial reaction-time tasks? *Psychological Research/Psychologische Forschung, 72*(4), 387-396.

Jimenez, L., Mendez, C., & Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(4), 948-969.

Jimenez, L., Vaquero, J. M. M., & Lupianez, J. (2006). Qualitative differences between implicit and explicit sequence learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 32*(3), 475-490.

Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1534-1555.

Kinder, A., Shanks, D. R., Cock, J., & Tunney, R. J. (2003). Recollection, fluency, and the explicit/Implicit distinction in artificial grammar learning. *Journal of Experimental Psychology: General, 132*(4), 551-565.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for specific instances. *Psychological Science, 3*(3), 172-179.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(1), 79-91.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 169-181.

Koch, I., & Hoffmann, J. (2000). Patterns, chunks, and hierarchies in serial reaction-time tasks. *Psychological Research, 63*(1), 22-35.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present day american english*. Providence: Brown university press.

Kuhn, G., & Dienes, Z. (2005). Implicit learning of nonlocal musical rules: Implicitly learning more than chunks. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(6), 1417-1432.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*, 294-340.

Lotz, A., & Kinder, A. (2006). Transfer in artificial grammar learning: The role of repetition information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 707-715.

Lozito, J. P., & Mulligan, N. W. (2010). Exploring the role of attention during implicit memory retrieval. *Journal of Memory and Language, 63*(3), 387-399.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30*(4), 607-613.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*(3), 252-271.

Marvel, C. L., Turner, B. M., O'Leary, D. S., Johnson, H. J., Pierson, R. K., Boles, L. L., & Andreasen, N. C. (2007). The neural correlates of implicit sequence learning in schizophrenia. *Neuropsychology, 21*(6), 761-777.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724-760.

Merikle, P. M., & Daneman, M. (1998). Psychological investigations of unconscious perception. *Journal of Consciousness Studies, 5*(1), 5-18.

Meulemans, T., & Van der Linden, M. (1997). Associative chunk strength in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 1007-1028.

Mickes, L., Johnson, E. M., & Wixted, J. T. (2010). Continuous recollection versus unitized familiarity in associative recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(4), 843-863.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*(5), 858-865.

Miller, G. A. (1958). Free recall of redundant strings of letters. *Journal of Experimental Psychology, 56*(6), 485-491.

Miyawaki, K. (2006). The influence of the response-stimulus interval on implicit and explicit learning of stimulus sequence. *Psychological Research-Psychologische Forschung, 70*(4), 262-272.

Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1095-1110.

Musen, G., & Squire, L. R. (1993). On the implicit learning of novel association by amnesic patients and normal subjects. *Neuropsychology, 7*(2), 119-135.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109-133.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*(1), 1-32.

Nissley, H. M., & Schmitter-Edgecombe, M. (2002). Perceptually based implicit learning in severe closed-head injury patients. *Neuropsychology, 16*(1), 111-122.

Norman, E., Price, M. C., Duff, S. C., & Mentzoni, R. A. (2007). Gradations of awareness in a modified sequence learning task. *Consciousness and Cognition: An International Journal, 16*(4), 809-837.

Onyper, S. V., Zhang, Y. X., & Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology: General, 139*(2), 341-364.

Overgaard, M., Timmermans, B., Sandberg, K., & Cleeremans, A. (2010). Optimizing subjective measures of consciousness. *Consciousness and Cognition: An International Journal, 19*(2), 682-684.

Ozubko, J. D., & Joordens, S. (2008). Super memory bros.: Going from mirror patterns to concordant patterns via similarity enhancements. *Memory & Cognition, 36*(8), 1391-1402.

Pacton, S., Fayol, M., & Perruchet, P. (2005). Children's implicit learning of graphotactic and morphological regularities. *Child Development, 76*(2), 324-339.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology-General, 130*(3), 401-426.

Pedersen, A., Siegmund, A., Ohrmann, P., Rist, F., Rothermundt, M., Suslow, T., & Arolt, V. (2008). Reduced implicit and explicit sequence learning in first-episode schizophrenia. *Neuropsychologia, 46*(1), 186-195.

Perlman, A., & Tzelgov, J. (2006). Interactions between encoding and retrieval in the domain of sequence-learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 32*(1), 118-130.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General, 119*(3), 264-275.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10*(2), 257-261. doi: 10.1038/nn1840

Persaud, N., McLeod, P., & Cowey, A. (2008). Commentary to note by Seth: Experiments show what post-decision wagering measures. *Consciousness and Cognition: An International Journal, 17*(3), 984-985. doi: 10.1016/j.concog.2007.06.002

Poznanski, Y., & Tzelgov, J. (2010). Modes of knowledge acquisition and retrieval in artificial grammar learning. *The Quarterly Journal of Experimental Psychology, 63*(8), 1495 - 1515.

Rauch, S. L., Wright, C. I., Savage, C. R., Martis, B., McMullin, K. G., Wedig, M. M., . . . Keuthen, N. J. (2007). Brain activation during implicit sequence learning in individuals with trichotillomania. *Psychiatry Research-Neuroimaging, 154*(3), 233-240.

Rausei, V., Makovski, T., & Jiang, Y. V. (2007). Attention dependency in implicit learning of repeated search context. *Quarterly Journal of Experimental Psychology, 60*(10), 1321-1328.

Reber, A. S. (1967). Implicit learning of artifical grammars. *Journal of Verbal Learning & Verbal Behavior, 6*(6), 855-863.

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology, 81*(1), 115-119.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*(3), 219-235.

Reber, P. J., & Kotovsky, K. (1997). Implicit learning in problem solving: The role of working memory capacity. *Journal of Experimental Psychology-General, 126*(2), 178-203.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A reexamination of stimulus-frequency effects in recognition: Two mirrors for low- and high-frequency pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 138-152.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 294-320.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests - Determining what is learned about sequence structure. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(3), 585-594.

Reed, J., & Johnson, P. (1998). Implicit learning: Methodological issues and evidence of unique characteristics. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning.* (pp. 261-294). Thousand Oaks, CA US: Sage Publications, Inc.

Reed, N., McLeod, P., & Dienes, Z. (2010). Implicit knowledge and motor skill: What people who know how to catch don't know. *Consciousness and Cognition: An International Journal, 19*(1), 63-76.

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics, 44*(6), 563-575.

Remillard, G. (2008). Implicit learning of second-, third-, and fourth-order adjacent and nonadjacent sequential dependencies. *The Quarterly Journal of Experimental Psychology, 61*(3), 400-424.

Remillard, G., & Clark, J. M. (2001). Implicit learning of first-, second-, and third-order transition probabilities. *Journal of Experimental Psychology-Learning Memory and Cognition, 27*(2), 483-498.

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 305-320.

Rowland, L. A., & Shanks, D. R. (2006). Sequence learning and selection difficulty. *Journal of Experimental Psychology-Human Perception and Performance, 32*(2), 287-299.

Runger, D., & Frensch, P. A. (2008). How incidental sequence learning creates reportable knowledge: The role of unexpected events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1011-1026.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1264-1288.

Scott, R. B., & Dienes, Z. (2010). Knowledge applied to new domains: The unconscious succeeds where the conscious fails. *Consciousness and Cognition: An International Journal, 19*(1), 391-398.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(4), 592-608.

Seth, A. K. (2008). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition: An International Journal, 17*(3), 981-983. doi: 10.1016/j.concog.2007.05.008

Shanks, D. R., Rowland, L. A., & Ranger, M. S. (2005). Attentional load and implicit sequence learning. *Psychological Research-Psychologische Forschung, 69*(5-6), 369-382.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*, 367-447.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179-195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145-166.

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37*(7), 976-984. doi: 10.3758/mc.37.7.976

Song, S., Howard, J. H., Jr., & Howard, D. V. (2007). Implicit probabilistic sequence learning is independent of explicit awareness. *Learning & Memory, 14*(3), 167-176.

Song, S., Howard, J. H., Jr., & Howard, D. V. (2008). Perceptual sequence learning in a serial reaction time task. *Experimental Brain Research, 189*(2), 145-158.

Stadler, M. A. (1995). Role of Attention Implicit Learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 21*(3), 674-685.

Stadler, M. A., & Roediger, H. L., III. (1998). The question of awareness in research on implicit learning. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning.* (pp. 105-132). Thousand Oaks, CA US: Sage Publications, Inc.

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63*(1), 18-34.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1379-1396.

Sun, R., Zhang, X., Slusarz, P., & Mathews, R. (2007). The interaction of implicit learning, explicit hypothesis testing learning and implicit-to-explicit knowledge extraction. *Neural Networks, 20*(1), 34-47.

Tamayo, R., & Frensch, P. A. (2007). Interference produces different forgetting rates for implicit and explicit knowledge. *Experimental Psychology, 54*(4), 304-310.

Tunney, R. J. (2005). Sources of confidence judgments in implicit cognition. *Psychonomic Bulletin & Review, 12*(2), 367-373.

Tunney, R. J. (2007). The subjective experience of remembering in artificial grammar learning. *European Journal of Cognitive Psychology, 19*(6), 934-952.

Tunney, R. J. (2010). Similarity and confidence in artificial grammar learning. *Experimental Psychology, 57*(2), 160-168.

Tunney, R. J., & Altmann, G. T. M. (2001). Two modes of transfer in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 614-639.

Tunney, R. J., & Bezzina, G. (2007). Effects of retention intervals on receiver operating characteristics in artificial grammar learning. *Acta Psychologica, 125*(1), 37-50.

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition, 31*(7), 1060-1071.

Turk-Browne, N. B., Junge, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology-General, 134*(4), 552-564.

Van den Bos, E., & Poletiek, F. H. (2008). Effects of grammar complexity on artificial grammar learning. *Memory & Cognition, 36*(6), 1122-1131.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*(2), 254-262.

Visser, T. A. W., & Merikle, P. M. (1999). Conscious and unconscious processes: The effects of motivation. *Consciousness and Cognition, 8*(1), 94-113.

Vokey, J. R., & Brooks, L. R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 328-344.

Vokey, J. R., & Higham, P. A. (2004). Opposition logic and neural network models in artificial grammar learning. *Consciousness and Cognition: An International Journal, 13*(3), 565-578.

Vokey, J. R., & Higham, P. A. (2005). Abstract analogies and positive transfer in artificial grammar learning. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Experimentale, 59*(1), 54-61.

Whittlesea, B. W. A., & Dorken, M. D. (1993). Incidentally, things in general are particularly determined - an episodic-processing account of implicit learning. *Journal of Experimental Psychology-General, 122*(2), 227-248.

Whittlesea, B. W. A., & Masson, M. E. J. (2005). Repetition blindness in rapid lists: Activation and inhibition versus construction and attribution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 54-67.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, Version 2.00. *Behavior Research Methods, Instruments & Computers, 20*(1), 6-10.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152-176.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*(4), 1025-1054. doi: 10.1037/a0020874

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*(4), 616-641.

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.