UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

School of Psychology

Fluency-Based Production and Memorability-Based Reduction of False Alarms in Recognition Memory

by

Helen Ho Yan Tam

Thesis for the degree of Doctor of Philosophy

August 2006

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES
SCHOOL OF PSYCHOLOGY

Doctor of Philosophy

FLUENCY-BASED PRODUCTION AND MEMORABILITY-BASED REDUCTION OF FALSE ALARMS IN RECOGNITION MEMORY

by Helen Ho Yan Tam

The production of false alarms in recognition memory tests has long been of interest to memory researchers. A recent paradigm devised to demonstrate false recognition was the "hension" effect paradigm (Whittlesea & Williams, 1998), where the false alarm (FA) rate for regular nonwords (e.g., HENSION) was found to exceed that for natural words (e.g., CURTAIN) and for irregular nonwords (e.g., STOFWUS).

The hension effect has been cited as empirical evidence for the discrepancy-attribution hypothesis, which assumes that the high FA rate for regular nonwords arose because the processing of these fluent, yet meaningless items is discrepant. Discrepancy in turn prompts fluency misattribution (i.e., false alarms) to occur.

An objective of this thesis was to examine the suitability of the discrepancy-attribution hypothesis in explaining the hension effect. In Experiments 1 – 4, the sense of discrepancy associated with the hension effect materials was manipulated. These experiments found that discrepancy did not appear to underlie false recognition.

As an alternative explanation for the hension effect, it was argued that recognition judgments are dependent on fluency-based processes for regular and irregular nonwords. However, the low FA rate observed for natural words was due to their high memorability levels (as substantiated by ratings data in Experiment 5), which allowed participants to correctly reject these items when they acted as lures.

Compelling evidence for the involvement of a memorability-based, metacognitive strategy in lure rejection came from the finding of a FA rate decrease for items whose memorability levels have been experimentally enhanced (Experiments 7 – 8). These results were interpreted from the perspective of two signal-detection models, one based on criterion shifts and one based on distribution shifts (a multi-process model). Support for the multi-process model was found in Experiments 9 – 10, where it was demonstrated that lure groups of differing intrinsic (item-based) and extrinsic (experimentally-manipulated) memorability levels are located on distinctly separate points on a hypothetical strength-of-evidence scale.

List of Contents

List of Tables and Figures

Tables

Figures

Acknowledgements

This thesis would not have come to its completion without the help and support of many people, both at the University of Southampton and back home in Australia. To begin, I am much indebted to my supervisor Dr Phil Higham for his advice and comments on this PhD project. I also thank Professors Nick Donnelly and Constantine Sedikides for their encouragements. In terms of technical assistance, I am deeply grateful to Luke Phillips, and in particular to Dr Dave Brook, without whose patient guidance on the programming of the Revolution software, none of the experiments reported in this thesis would have existed. I am also indebted, quite literally, to the School of Psychology and the Overseas Research Student Award Scheme, for giving me this wonderful opportunity to study in Southampton.

Every aspect of this thesis has been shaped by many others, all for the better. I send my heartfelt thanks to: my family – Maria, Francis, Henry, my friends – K., Jeta, Yami, Angelo, Jenny, ME and Roberta, my officemates over the past years – Mark, Xingshan, Cara, Ina, Ailsa, and Tamas (who all demonstrated saintly forbearance in light of the negative vibes emanating from my corner), the people who have provided the soundtrack – Ben, Roy & HG, John Safran & Father Bob Macguire, Tony Delroy, and everyone else who have been so kind in wishing me well during my time here.

Finally, I express my sincere gratitude to all my participants, who have provided me with so much interesting data, and therefore so much to write about in this thesis.

Abbreviations

ALT	Attention-likelihood theory
ANOVA	Analysis of Variance
C	Bias estimate
cm	centimetres
d'	*d* prime
DRM	Deese-Roediger-Dermott (false memory paradigm)
F	F ratio
FA	False alarm
ITI	Inter-trial interval
LDT	Lexical decision task
M	Mean
MSE	Mean square error
ms	Milliseconds
N	Sample size
n	Sample size (subgroup)
p	probability
PDP	Process dissociation procedure (Jacoby, 1991)
REM	Retrieving Effectively from Memory (Shiffrin & Steyvers, 1997)
s	Seconds
SCAPE	Selective Construction and Preservation of Experiences (Whittlesea, 1997)
SD	Signal-detection
SE	Standard error
t	*t* statistic
WFE	Word frequency effect
η^2	Partial eta squared

Preface

The fallibility of human memory has been the subject of investigation since the very beginnings of experimental psychology. One aspect of memory failure, that of forgetting, was first examined in the pioneering work of Ebbinghaus in the late 19th century, and continues to be a subject of interest in more recent studies on phenomena such as retrieval-induced-forgetting (e.g., Anderson, Bjork, & Bjork, 1994) and directed forgetting (e.g., Bjork, 1989; Sahakyan & Delaney, 2003). In contrast, the other failure of memory, that of illusory memories for events that have not actually occurred, has a somewhat fragmented history. Arguably, the first investigation on this subject was carried out in the early 20th century by Alfred Binet in his systematic examination of unreliable recollection in children (Roediger & McDermott, 2000). Later, Bartlett (1932) also demonstrated incidences of false memories when he conceptualised memory as a reconstructive process, but one which was prone to schema-influenced errors. The study of false memories went through the doldrums in the mid 20th century, when errors in memory tests were actively factored out because they were merely seen as noise which masked true memory performance. However, following the cognitive revolution in the 1960s, research on memory illusions was revived through the work of researchers such as Bransford and Franks (1973) and Loftus (1979). Since then, the topic of false memories has formed an integral part of memory research. There are several reasons for this. As exemplified by much of Loftus's work, our understanding of the nature of false memories have important social implications in the legal domain - the credibility of eyewitness testimony and the veracity of apparently recovered memories (e.g., those concerning childhood abuse) remain a contentious issue. Away from the applied perspective, for memory theorists, false memories are in themselves intriguing phenomena which are worthy of scientific investigation. Indeed some researchers have argued that in the same way perceptual illusions have contributed to our understanding of the visual system, the study of memory illusions may provide us with valuable insights into the true workings of human memory (Roediger & McDermott, 2000).

One context in which false memories could occur is in recognition, whereby a novel, never-before-seen stimulus is falsely claimed to have been encountered previously. Instances of false recognition are found in everyday life – when a crime

victim falsely accuses an innocent suspect, or when a distractor (false option) is chosen as the correct answer in a multiple-choice test. Experimentally, false memories (and general performance) in recognition are examined through recognition tests devised by the experimenter. The format of a typical recognition test consists of a study phase, followed by a test phase. During the study phase, participants are presented with a set of items to be committed to memory. In the test phase, participants are presented with a mixture of targets (i.e., "old", or studied items) and lures (i.e., "new", or non-studied items). The participants' task is to discriminate the targets from the lures. In other words, correct responding entails saying "old" to a studied item (also known as "hit") and "new" to a non-studied item ("correct rejection"). Incorrect responding, on the other hand, entails judging a studied item as "new" ("miss") and a non-studied item as "old" ("false alarm", or FA). It can be seen that hits and misses, and false alarms and correct rejections, are both complementary pairs in that the two measures of the pair add up to the number of targets and lures respectively. Hence, a participant's recognition performance can be sufficiently indexed by the hit rate (proportion of targets correctly judged as old) and FA rate (proportion of lures incorrectly judged as old).

To investigate the phenomenon of false recognition, psychologists devise specific paradigms which would predict and produce noteworthy patterns of FA rates. An example of such a paradigm can be found in Whittlesea and Williams (1998). These researchers created a recognition test consisting of three distinctly different types of items – (a) natural words (i.e., real English words, e.g., CURTAIN), (b) regular nonwords (i.e., nonsense words which are easy to pronounce, e.g., HENSION), and (c) irregular nonwords (i.e., nonsense words which are difficult to pronounce, e.g., STOFWUS). Using these items, these researchers produced a remarkably robust finding whereby the rate of false recognition (as indexed by the FA rate) was greater for regular nonwords than for natural words or irregular nonwords. Whittlesea and Williams dubbed this pattern of FA rates "the hension effect", after one of the regular nonword items.

On the basis of their findings, the *discrepancy-attribution hypothesis* was put forward by Whittlesea and Williams (1998) as a putative account for how false recognition occurs. In subsequent work by Whittlesea and his colleagues, new

concepts and experimental manipulations were further introduced to gather supporting evidence for this hypothesis, and to generalise this principle to other illusions found in different memory paradigms and to other aspects of human judgments and cognition (e.g., Whittlesea, 2002a, 2002b; Whittlesea & Leboe, 2000, 2003; Whittlesea, Masson, & Hughes, 2005; Whittlesea & Williams, 2000, 2001a, 2001b). Elsewhere, little research has emerged to examine this hypothesis, or perhaps more fundamentally, the validity of using this hypothesis to explain the hension effect. Thus, one of the aims of this thesis is to appraise the suitability of the discrepancy-attribution hypothesis in accounting for the hension effect. In the first part of the thesis, a background of the research and theories on the causes of false recognition will be presented, along with a detailed exposition of Whittlesea and Williams's framework. This will be followed by the report of several experiments. These experiments were based on the hension effect paradigm, and were conducted to evaluate some of the predictions derived from the discrepancy-attribution hypothesis. It is important to note that the goal of this thesis is *not* to refute the hypothesis itself, but rather to scrutinise whether the hypothesis is the appropriate explanation for the hension effect, and in so doing, to investigate whether alternative factors may play a role in producing the effect. To this end, the latter parts of the thesis will describe other experiments which were carried out to address this issue. Specifically, it will be argued that metacognitive processes, in particular those based on the assessment of item memorability, may be an influential factor in the production of false alarms, and therefore should be incorporated in the development of recognition memory models.

Chapter 1

Introduction

> "And it is an assumption made by many writers that the revival of an image is all that is needed to constitute the memory of the original occurrence. But such a revival is obviously not a *memory*, whatever else it may be; it is simply a duplicate...A farther condition is required before the present image can be held to stand for a *past original*."
>
> "That condition is that the fact imaged be *expressly referred to the past*, thought as *in the past*."
>
> - William James (1890)

1.1 Feelings of Remembering are Products of Attributions

In his seminal work, *Principles of Psychology*, William James alluded to the notion that in order to fully understand how memory works, we need to explain how a rememberer comes to evaluate a "revival of an image" in memory as a representation of an original event in the past. Nearly a century later, these sentiments were echoed by Jacoby, Kelley and Dywan (1989) in their paper titled "Memory Attributions", where it was proposed that subjective feelings of remembering[1], as experienced by, say, a participant in a memory task, are "an attribution of a response to a particular cause; that is, to the past." (p. 391). In this, the important role played by subjective experience in remembering was emphasised by Jacoby et al., who argued that it is this subjective experience which drives behaviour. As an example, they noted that amnesics could have intact procedural memory, but yet are reluctant to act upon this retrieved information in the absence of a subjective experience of remembering. Thus,

[1] The term remembering is used here as a generic term to refer to instances where the participant has grounds to claim that something has happened in the past. Therefore, it should not be confused here with the identical-in-name concept in the remember/know distinction advocated by Tulving (1985), where "remember" and "know" responses are generally equated with recollection- and familiarity-based processes respectively (e.g., Gardiner, 1988). These concepts, along with the prominent dual-route model of recognition memory (e.g., Jacoby & Dallas, 1981; Mandler, 1979, 1980), will be elaborated later in this chapter.

the phenomenology of remembering is crucial in giving one "the impetus to act" (Jacoby et al., p. 400).

The novel contribution of Jacoby et al.'s (1989) work, however, lies in their speculation that attributional mechanisms could underlie not only veridical recognition, but also false recognition. They noted that attributional accounts, as seen elsewhere in the research on emotions, suggests a less-than-direct relationship between physiological responses and the nature of emotive feelings one experiences. Depending on one's situational context, adrenalin-induced arousal could be interpreted as anger or happiness (Schachter & Singer, 1962). In a similar way, Jacoby et al. proposed that the strength (or even the presence) of memory representations might not have a direct causal relationship with feelings of remembering. Certainly, a lack of subjective experience of remembering could reflect an absence of that particular episode being represented in memory, or that the representation had failed to be retrieved or activated. Conversely, subjective feelings of remembering indicate the presence of such a trace in memory. However, observations on amnesic patients, as well as on normal subjects, suggest that the link between memory representation and subjective feelings of remembering may not be as straightforward as is commonly believed (Jacoby & Witherspoon, 1982). Specifically, Jacoby et al. argued that feelings of remembering may not necessarily result from a re-activation of a representational trace in memory, but rather are products of an "attribution or unconscious inference" (p. 392). False recognition, therefore, would occur when attributional processes produces phenomenological experiences of pastness, in the absence of an actual memory trace of the "recognised" event.

1.2 The Basis of the Memory Attributions - Fluency

In Jacoby et al.'s (1989) account, *fluency*, as experienced during the processing for a given stimulus, is the crucial element which underpins memory attributions. The idea of fluency had emerged as early as the turn of the last century, when Titchener described a fluent stimulus as a "reconstruction along the line of least resistance" (cited in Jacoby et al., 1989, p. 395). The phenomenology associated with fluency was also illustrated by Baddeley's (1982) description that fluently processed stimuli seem to "pop into mind". Likewise, Jacoby et al. (1989) identified fluency

with processing ease (e.g., "words read once are more easily re-read later", p. 401), and the speed of processing (e.g., "an idea considered once comes to mind more readily later", p. 401). Based on these ideas, the operational definition commonly held by memory researchers is that fluency can be measured in terms of the speed (response latency, e.g., time taken to pronounce a word stimulus, see Whittlesea, Jacoby, & Girard, 1990) and/or the accuracy by which a stimulus is identified (e.g., Jacoby & Dallas, 1981). The principal conjecture put forward by Jacoby et al. was that fluency experienced during the current processing of a stimulus could be interpreted as an indicator that this stimulus had been encountered in the past.

Much of the groundwork underlying Jacoby et al's (1989) hypothesis that subjective feelings of remembering arise from processing fluency could be found in Jacoby and Dallas's (1981) investigation on repetition (or perceptual) priming. In their investigation, repetition priming referred to the way that accuracy in identifying a briefly flashed stimulus (presented for a duration of 35 ms) could be greatly enhanced from prior study of that stimulus. A significant finding from Jacoby and Dallas was that the effects of repetition priming were more robust than first thought – perceptual identification performance was found to benefit even from a brief previous study (1 s) of the stimulus, and this enhancement was retained despite a 24-hour delay between study and test. Importantly, however, Jacoby and Dallas extended their investigation by examining the relationship between repetition priming and recognition memory. They argued that the speed and accuracy by which perceptual identification was performed could be used by participants in discriminating studied (old) from not-studied (new) items in a recognition test. Because previously encountered stimuli are processed more fluently (i.e., due to repetition priming), processing fluency experienced for a given stimulus could therefore be attributed to its prior occurrence.

1.3 Fluency Misattribution as the Cause of False Recognition

It can be seen then that fluency in processing could be a reliable indicator for veridical recognition. That is, feelings of remembering produced through fluency attributions do reflect the reality that the stimulus had actually occurred in the past. Critically, however, Jacoby et al. (1989) postulated that the same fluency-based attributional process could underlie instances of false recognition. These theorists

argued that processing fluency for a stimulus could be boosted by *factors other than prior experience*. In other words, if the processing of a never-before-seen stimulus is, for some reason, fluent, then this kind of fluency could subsequently be *misattributed* to a source in the past, thus resulting in false recognition.

A demonstration of this type of illusory recognition can be found in Jacoby and Whitehouse (1989), who devised a recognition test where each test word item was preceded by a briefly presented prime (the presentation duration was 16 ms). The prime took one of three forms: (a) matching – a word which was identical to the test word, (b) nonmatching – a word which was different to the test word, or (c) control – a series of letters ("xoxoxox"). Jacoby and Whitehouse found that significantly more "old" judgments were given in the matching prime, than in the nonmatching or control prime conditions, regardless of the test item's actual old/new status. These researchers attributed this finding to the way that a matching prime would enhance the processing fluency of the subsequent test word, and in turn, the fluency produced would be interpreted by participants to be an indicator of pastness.

Similarly, Whittlesea et al. (1990) developed a paradigm to examine fluency-based false recognition. On a given trial in their experiment, participants were presented with a succession of seven words, each displayed briefly for 60 ms. This serial presentation of seven words was followed immediately by the test word which was occluded either by a light mask or a heavy mask. The difference between the two levels of mask density used was discernible, but not obvious. For the test word, participants were required to first pronounce it, and then judge the test word as old if it was one of the seven preceding words that were just presented, or new if it was not one of those seven words. Analyses on participants' pronunciation latencies revealed two significant outcomes. First, test words that were old were pronounced more quickly than test words that were new. This finding was consistent with the notion of repetition priming, because identification/pronunciation of the test word was enhanced by its previous presentation. The second finding from the pronunciation latency data was that unsurprisingly, test words presented under a light mask were pronounced more quickly than those presented under a heavy mask. Importantly, this result was accompanied by recognition data showing that test words, regardless of their actual old/new status, were more likely to be judged as old if they were

presented under the light rather than the heavy mask. Thus, in the case of a new test word, processing fluency that arose because of the light mask condition was being misattributed to the stimulus's prior occurrence (Whittlesea et al., 1990).

1.4 Bidirectional Influence Between the Current Processing of a Stimulus and its Prior Occurrence

Additional experiments conducted by Whittlesea et al. (1990) further showed that the relationship between processing fluency and memory appeared to be bidirectional. Fluency resulted from an actual past encounter with the stimulus could affect the current perception of the stimulus. In one experiment, Whittlesea et al. utilised the same experimental paradigm as described above, but imposed a different task demand on the participants. Rather than making an old/new judgment, participants were required to assess the density of the mask occluding the presented test word. In this case, processing fluency due to the test word being old resulted in the mask being perceived as less dense.

This last finding by Whittlesea et al. (1990) was predated by a large body of research on how memory could influence the perception of later events. For example, when participants were asked to judge the length of presentation duration for test items, items that were previously studied were judged to be presented for a longer duration than those that were not studied, regardless of whether the duration was actually long or short (Witherspoon & Allen, 1985). Likewise, in a task where participants were to judge the noise level within which sentences were embedded, the noise level was perceived to be less loud when a previously studied, rather than a new sentence, was presented concurrently with the noise (Jacoby, Allan, Collins, & Larwill, 1988). Similarly, previously encountered geometric shapes were subsequently judged to be brighter, or darker (depending on the task specification), than new stimuli (Mandler, Nakamura, & Van Zandt, 1987).

Further, not only could the effect of prior experience be manifested in the perception of the stimulus's physical characteristics (e.g., clarity, duration, loudness, brightness, etc.), it could also modify the perception of the stimulus's non-physical attributes (e.g., meaning, pleasantness). A prior encounter of a nonfamous name could render it to be perceived as famous in a subsequent fame judgment task (Jacoby,

Kelley, Brown, & Jasechko, 1989; Jacoby, Woloshyn, & Kelley, 1989); previously presented geometrical shapes were preferred to new ones (Mandler et al., 1987); and words that had been studied were judged to be more pleasant than those that had not (Whittlesea, 1993).

Collectively, studies by Whittlesea et al. (1990) and others (e.g., Jacoby et al., 1988; Mandler et al. 1987) suggest that the dimension of the stimulus to which fluency is attributed may be largely dependent on task demands. Fluent processing actually caused by past experience with the stimulus can be misinterpreted as a product of a physical or non-physical characteristic inherent in the presently encountered stimulus. Conversely, fluent processing actually caused by a current perceptual manipulation can be misattributed to the past, resulting in false recognition.

1.5 Dissociation Between Fluency and Recognition Memory Judgments

Thus far, it appears that in a recognition task, processing fluency would, through attributional mechanisms, be invariably translated to "old" judgments. The relationship between fluency and feelings of remembering, however, is in fact far more complex. Indeed, while Jacoby and Dallas (1981) noted the robustness of repetition priming, they also demonstrated that processing fluency (in this case, measured by the identification accuracy of briefly presented test stimuli) could be dissociated from recognition memory performance. More specifically, these researchers found that certain variables could affect recognition memory performance without necessarily affecting perceptual identification success. For instance, in one experiment, participants encoded some items using deep processing which required the elaboration of the items' semantic properties, and encoded other items using shallow processing which only focused on the items' physical characteristics. As predicted from the level-of-processing account (Craik & Lockhart, 1972), recognition (in this case, indexed by hit rate) was better for items that underwent deep, semantic encoding than for items that were only shallowly encoded. In contrast, accuracy in perceptual identification remained constant despite of this level-of-processing manipulation. Likewise, imposing a 24-hour delay between study and test substantially impaired recognition memory while leaving perceptual identification relatively unaffected (Jacoby & Dallas, 1981). Such dissociation between recognition

memory and perceptual identification performance suggests that while some studies (e.g., Whittlesea et al., 1990) indicated a direct link between processing fluency and recognition judgments, this relationship might not be a straightforward one.

The observation of a low correlation between fluency and the likelihood of an "old" judgment has also been noted by Johnston, Dark, and Jacoby (1985). These authors argued that if recognition memory is solely dependent on processing fluency (indexed here by pronunciation latency), then items more fluently processed (with a faster pronunciation latency) would elicit an "old" response (i.e., produce hits and false alarms), whereas items less fluently processed (with a slower pronunciation latency) would elicit a "new" response (i.e., produce misses and correct rejections). Consistent with this account, the average pronunciation latency was indeed faster for hits than for correct rejections. However, contrary to their prediction, Johnston et al. found that the average pronunciation latency was actually faster for misses than for false alarms. That is, despite the fact that old items were processed more fluently than new items in a recognition test, some of these old items were mistakenly judged to be new. Conversely, some of the new items were incorrectly recognised as old, despite the fact that processing fluency for these items was low (Johnston et al., 1985; see also Poldrack & Logan, 1998 for similar evidence that fluency, or speed of item processing mediates recognition performance only to a limited extent).

1.6 The Dual-Process Model of Recognition Memory

To account for the indirect relationship between processing fluency and recognition judgments (e.g., Jacoby & Dallas, 1981; Johnson et al., 1985), Jacoby and his colleagues advocated a dual-process account of recognition memory. First originated in the 1970's (e.g., Atkinson & Juola, 1973, 1974), the dual-process theory has become a prominent force in recognition memory, inspiring the proposal of a large number of memory models in the past 30 years (e.g., Atkinson & Juola, 1973; Huppert & Piercy, 1976; Mandler, 1979, 1980; Jacoby, 1991; Tulving, 1982, Tulving & Schacter, 1990; Yonelinas, 1994, 1997, 1999, 2001; see Yonelinas, 2002 for a review). Despite some critical differences among them, the basic assumption shared by these models is that recognition memory is based on two distinct processes, recollection and familiarity. A common conception of these two processes, as proposed by Jacoby and his colleagues (e.g., Jacoby & Dallas, 1981, Jacoby, Kelley,

& Dywan, 1989, Jacoby & Witherspoon, 1982), is that recollection entails the recovery of the context in which the recognised stimulus was originally encoded, whereas a familiarity-based recognition judgment involves the assessment of processing fluency of the encountered stimulus. Thus, according to this particular class of dual-process models, when processing fluency of a stimulus is attributed to its pastness, the phenomenology of remembrance is specified to be feelings of *familiarity* for that stimulus. Moreover, in agreement with other models such as those proposed by Atkinson and Juola, and Mandler, Jacoby's dual-process account assumes that recollection- and familiarity-related mechanisms operate independently and in parallel. Additionally, while recollection is argued to be an effortful and controlled process, familiarity, on the other hand, is described as a quick and largely automatic route to recognition (Yonelinas, 2002)[2].

From the perspective of the dual-process model, the low correlation between fluency and "old" recognition judgments could be ascribed to the involvement of recollection-based mechanisms in recognition. If recognition memory is solely based on processing fluency, variables affecting recognition memory performance would in turn have an impact on perceptual identification performance. The dissociation found by Jacoby and Dallas (1981) whereby identification accuracy remained stable, while recognition memory performance fluctuated, runs against the notion that the probability of a stimulus being judged as old could be predicted directly from its processing fluency. These researchers argued that, in the example where recognition was found to benefit from deep, semantic processing, participants were basing their recognition judgments on the recollection of study-context details, rather than on the item's processing fluency. Likewise, the fact that processing fluency was found to be greater for old items judged to be new (misses) than for new items judged to be old (false alarms) suggests a reliance on a separate recollection-based process that was not contingent on fluency (Johnston et al., 1985). However, because this recollection route was argued to be "fallible" (Johnston et al., 1985, p. 7), reliance on this route would therefore render some old items to be erroneously judged as new. Presumably, this would occur when despite the studied item's high levels of processing fluency,

[2] However, more recent reaction-time data suggested that recollection-based responses are made faster than familiarity-based responses (Dewhurst, Holmes, Brandt, & Dean, 2006; Wixted & Stretch, 2004).

the participant fails to retrieve sufficient recollective details for it to warrant an "old" judgment.

Thus far, the assumptions of the dual-process model delineated here primarily concern veridical recognition, that is, according to the model, the participants' ability to produce hits on a recognition test, is based on either recollection or familiarity. In terms of false recognition, the model assumes that false alarms are driven by the familiarity process. Indeed, the process dissociation procedure (PDP), which was devised by Jacoby (1991, 1999; see also Hay & Jacoby, 1996, 1999) to investigate the role of recollection and familiarity in recognition, hinges on these above assumptions on how hits and false alarms are produced. To give an example, in one version of the PDP, participants were first shown a list of visually presented word, followed then by a list of auditorily presented word which they were required to learn intentionally for a later test (Jacoby, 1991). Two different tests were subsequently given to participants – an inclusion and an exclusion test. In the inclusion test, participants were asked to respond "yes" to items that were either seen (in the first list) or heard (in the second list). In the exclusion test, they were to respond "yes" only to items that were heard, and "no" to items that were seen. The rationale underlying the PDP is that both recollection and familiarity could be used in responding to items in the inclusion test, whereas only recollection of contextual details associated with encoding would allow participants to discriminate heard items from seen items in the exclusion test. Moreover, seen items that were incorrectly claimed to be heard in the exclusion test were assumed to be a reflection of the influence of familiarity on recognition (Jacoby, 1991)[3].

[3] Mathematically, these above assumptions could be expressed by the following equations (Jacoby, 1991). In the inclusion test, the probability of a "yes" response is equal to the probability of the item being recollected *and* the probability of the item being accepted on the basis of familiarity (F), in the absence of recollection (R): P(inclusion) = R + F(1-R). In the exclusion test, the probability of incorrectly responding "yes" to a seen item is simply equal to the probability that the item is not recollected, but is familiar: P(exclusion) = F(1-R). From these equations, estimates of recollection and familiarity could be computed (Jacoby, 1991). Subtracting P(exclusion) from P(inclusion) would give an estimate of recollection, i.e., R = P(exclusion) – P(inclusion). An estimate of familiarity is calculated by dividing P(exclusion) by (1-R), i.e., F = P(exclusion)/(1-R).

Further support for the notion that false recognition is derived from a familiarity-based process could be found in a number of studies showing that manipulation on processing fluency, which purportedly underpins the familiarity process, would induce changes in FA rates, as well as changes in the production of "yes" responses associated with familiarity. For example, Rajaram (1993) incorporated remember/know response instructions (e.g., Gardiner, 1988; Gardiner, Ramponi, & Richardson-Klavehn, 2002) in order to elucidate the specific basis by which participants made "old" judgments in the Jacoby-Whitehouse paradigm (1989), as detailed earlier in this chapter. First developed by Tulving (e.g., 1982, 1985), the remember/know distinction aligns closely with the principal ideas espoused by the dual-process theory, in that "remembering" is associated with the recollection of details relating to the study episode, and "knowing" is associated with feelings of familiarity in the absence of recollection. In requiring participants to classify their "old" recognition judgments as either "remember" or "know", Rajaram showed that the increase in "old" judgments found in Jacoby and Whitehouse's matching prime condition was primarily driven by an increase in "know" responses. In contrast, the proportion of "old" judgments classified as "remember" was unaffected by the priming manipulation.

Elsewhere, as noted by Yonelinas (2002), recognition tests using the R/K response instructions generally produce very low levels of "R" responses for new items, with the majority of false alarms being classified as "K" (a selection of examples include Gregg & Gardiner, 1994; Guttentag & Carroll, 1997; Rajaram, 1993; Yonelinas, 2001; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). The only findings that are inconsistent to this trend in R/K responses can be found when the Deese-Roediger-McDermott (DRM) paradigm was used to produce false recognition (Roediger & McDermott, 1995; see also Read, 1996). In the DRM paradigm, participants are given study items (e.g., *bed*, *blanket*, *night*) which are all related to a critical theme word (e.g., *sleep*). The theme word is not shown during study, but acts as a critical lure at test. The rate of false recognition for the critical lure has been found to be remarkably high, even approximating the hit rate for related items that had actually been studied (e.g., Roediger & McDermott, 1995; Whittlesea, 2002a). Moreover, the majority of false alarms associated with the critical lure were classified as R, rather than K, thus suggesting that participants "recollected" details of

items that were not studied (e.g., Gallo, Roediger, & McDermott, 2001; Israel & Schacter, 1997; Schacter, Verfaellie, & Anas, 1997; Schacter, Verfaellie, & Pradere, 1996). However, alternative explanations have been offered for this contradictory finding (Schacter, Norman, & Koutstaal, 1998). For example, Norman and Schacter (1997) showed that information retrieved at test in a DRM experiment primarily pertained to semantic associations rather than sensory contextual details encoded during the study episode. False recollection would therefore arise because studied (related) items and the critical lure share a large number of semantic associations. Another possibility is that the critical lure is covertly generated by participants during study and this self-generated representation of the critical lure is subsequently recollected at test (e.g., Roediger & McDermott, 1995; see also Yonelinas, 2002). Thus, false recollection here constitutes a source-monitoring error (e.g., Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981) because the participant fails to discriminate between an actual experienced event from an imagined one.

Overall, the evidence from the DRM literature suggests that false recognition could arise from other psychological mechanisms in which the concept of processing fluency is not necessarily implicated. However, it might be important to note that hypothesised accounts like those described above (e.g., Schacter et al., 1998) are specific to findings from the DRM paradigm. In a standard recognition test consisting of word stimuli, experimenters are usually at pains to minimise the overlaps in semantic associations among items, and it is impossible for the participants to self-generate the lures devised by the experimenter. Thus, for the standard recognition test based on individual word stimuli, the general consensus among adherents of the dual-process model is that false alarms are produced through the familiarity route in recognition, when fluency experienced for a new test item is misattributed to the past.

1.7 The SCAPE Framework and the Discrepancy-Attribution Hypothesis

An alternative to the dual-route account in explaining the complex relationship between processing fluency and recognition memory can be found in the discrepancy-attribution hypothesis, as put forward by Whittlesea and Williams (1998, 2000, 2001a, 2001b). In their work, Whittlesea and Williams reported primarily on the conditions whereby fluency from current processing would be misattributed to the past, thus producing a false recognition. In particular, they were interested in how

under certain experimental manipulations, FA rates would exceed the level that would be predicted by fluency measures (e.g., pronunciation latency). The central point of their theory is that in order for such an augmented rate of misattribution to occur, participants need to experience a level of fluency that is surprising, or discrepant to their expectations.

The discrepancy-attribution hypothesis has its foundations in the Selective Construction And Preservation of Experiences (SCAPE) framework (Whittlesea, 1997, 2003). As a functional account of memory, the SCAPE framework assumes that there is one memory system responsible for the preservation of experiences and the control of behaviours. Moreover, this framework places an emphasis on the idea that memories are not "retrieved", but are reconstructed through two stages: production and evaluation. Production here refers to the performance of a cognitive operation on an encountered stimulus. The result of this production could be either a behavioural, or in the case of remembering, a mental event. Further, production is said to be influenced by the rememberer's attitudes and expectations towards this encountered stimulus. Following production, evaluation allows the rememberer to assess the way the cognitive operation has been performed on the stimulus. Whittlesea argued that it is at this evaluative stage, that in the words of Marcel (1983), participants are attempting to "make sense of as much data as possible at the most functionally useful level" (p. 238). In relation to memory, the subjective experience of remembering is therefore said to arise through this evaluative process (Whittlesea, 1997, 2003).

Unlike the dual-route account of recognition memory, the SCAPE framework does not assume recognition judgments which are perceived by participants to be "recollection-based" versus those which are "familiarity-based" necessarily correspond to separate, distinctive processes (for a similar view, see Gruppuso, Lindsay, & Kelley, 1997; and Higham & Vokey, 2004). Instead, the primary focus of the framework is to explain the subjective feelings of remembering one could experience when encountering a particular stimulus. Straying from the recollection/familiarity distinction, the terminology used in the SCAPE framework for these subjective feelings is "feelings of familiarity". These feelings, argued to be the *sine qua non* of the phenomenology of remembering (Whittlesea, 1993), are proposed to be the result of an inferential process that constitutes evaluation. It is also assumed

that inferences made during evaluation could be based on certain heuristics the rememberer is using (Whittlesea, 1997, 2003). In this sense, recognition judgments are made much in the same way some decisions are argued to be made in regards to the nature of our environment (e.g., Kahneman & Tversky, 1973; Tversky & Kahneman, 1971, 1973). For instance, decisions pertaining to the size of a category are proposed to be made with the use of the availability heuristic (Tversky & Kahneman, 1971), a rule of thumb which assumes the ease by which a category member comes to mind reflects the size of the category to which the member belongs. In a similar vein, recognition judgments could be performed on the basis of a "fluency heuristic" whereby processing fluency of a particular stimulus is deemed to be indicative of the stimulus's pastness.

1.8 Evidence for the Discrepancy-Attribution Hypothesis: The Hension Effect

The evaluative stage postulated in the SCAPE framework, however, suggests that the fluency heuristic would only be used in certain situations. Thus, it is not simply the case that processing fluency would unfailingly produce feelings of familiarity at every turn, and hence the relationship between processing fluency and recognition memory is not direct. Specifically, Whittlesea and Williams (1998, 2000, 2001a, 2001b) argued that the possibility of fluency-based attributions being made is largely dependent upon one's expectations and the evaluation of one's interaction with an encountered stimulus. This tenet, which lies at the heart of the discrepancy-attribution hypothesis, is supported by a series of experimental findings reported by these researchers. For example, in one experiment (Whittlesea & Williams, 1998, Experiment 1), participants studied a list containing natural words (e.g., BOTTLE), pseudohomophones (nonwords that sound identical to a real word when pronounced, e.g., PHRAUG), and nonwords (that are not pseudohomophones, e.g., BELINT). In a subsequent recognition test, participants were first required to pronounce each test item aloud, and then make an old/new judgment for the test item. It was found that for new items, pronunciation latencies were faster for natural words than pseudohomophones, which were in turn pronounced more quickly than nonwords. However, in terms of the probability of a new item being incorrectly recognised, the FA rate was significantly higher for pseudohomophones (43%) than for natural words (19%) or nonwords (14%, Whittlesea & Williams, 1998).

In another study, Whittlesea and Williams (1998, Experiment 2) devised three types of items: natural words (e.g., CURTAIN), regular nonwords (nonsense words that are easy to pronounce, e.g., HENSION), and irregular nonwords (nonsense words that are difficult to pronounce, e.g., STOFWUS). Participants were first presented with a list containing these three types of items, and in a subsequent recognition test, were asked to pronounce the test item, then perform a lexical decision task (i.e., decide whether the test item was a word or nonword), and finally judge whether the item was old or new. For new items, pronunciation and lexical decision latencies were fastest for natural words, followed by regular nonwords, with irregular nonwords being the slowest. FA rates, however, showed that substantially more regular nonwords (37%) than both natural words (16%) and irregular nonwords (9%) were incorrectly judged as old by participants. This effect in the FA rates was dubbed "the hension effect" by Whittlesea and Williams, after a regular nonword exemplar.

Consistent with findings provided by proponents of the dual-process theory (e.g., Jacoby & Dallas, 1981; Johnston et al., 1985), the results from Whittlesea and Williams (1998) effectively demonstrated the indirect relationship between processing fluency and judgments of old in recognition. However, unlike dual-process theorists who invoked the concept of recollection to account for the imperfect correlation between fluency and "old" judgments, Whittlesea and Williams's exposition focussed on the way processing fluency would be evaluated by the participant. They noted that natural words were pronounced more quickly than both pseudohomophones and regular nonwords. Hence, if feelings of oldness, as measured by the probability of "old" judgments, were produced directly from processing fluency, then one would expect more false alarms to be made for natural words than either pseudohomophones or regular nonwords. The reversed pattern of findings, which show a higher FA rate for pseudohomophones and regular nonwords than for natural words, therefore suggest that the fluency heuristic was being selectively applied by participants in their recognition decisions. Specifically, Whittlesea and Williams argued that the heuristic was applied for pseudohomophones and regular nonwords because the processing fluency for these items was perceived to be *surprising*. For pseudohomophones (e.g., PHRAUG), this surprise is derived from the initial expectation that these nonwords are meaningless and therefore would not be processed fluently, but the act of pronouncing these items would produce the realisation that their phonological

representations correspond to those for meaningful natural words (e.g., FROG). For regular nonwords (e.g., HENSION), it was argued that these items could be fluently processed, but with this fluency one would also expect that there would be corresponding meanings attached to these items. The realisation that these items are ultimately meaningless therefore violates this expectation. According to the discrepancy-attribution hypothesis, it is this discrepancy between one's initial expectation and one's subsequent actual experience with the stimulus which crucially predicts whether fluency misattribution would take place.

The discrepancy inherent in the processing of pseudohomophones and regular nonwords could be contrasted with the processing of other stimuli used by Whittlesea and Williams (1998). Although natural words were pronounced more quickly than pseudohomophones and regular nonwords, the processing fluency of these natural words was assumed to be unsurprising, as one would expect fluently pronounced items would also be meaningful (which they were). Similarly, the lack of processing fluency for non-pseudohomophone (BELINT) and irregular nonwords (STOFWUS) was unsurprising to participants because both of these item types were meaningless and the lack of processing fluency experienced for these items was therefore consistent with the items' meaninglessness. Thus, compared to the FA rate yielded for pseudohomophones and regular nonwords, FA rates were lower for natural words, non-pseudohomophones and irregular nonwords because there was no discrepancy between the expectation and the actual processing experienced for these items. On the other hand, for pseudohomophones and regular nonwords, it was proposed that their level of subsequent processing fluency violated the initial expectations formed for these items, thus creating a sense of surprise and discrepancy. It was argued that in the context of a recognition task, this sense of surprise would be attributed to a prior encounter of these items (Whittlesea & Williams, 1998).

1.9 Feelings of Familiarity

In contrast to the dual-process model, the discrepancy-attribution hypothesis formulated by Whittlesea and Williams (1998, 2000, 2001a, 2001b) has its emphasis on false recognition, and particularly in a specific type of "feeling of familiarity" which was described as "the subjective feeling of having prior experience, whether or not one actually has" (Whittlesea & Williams, 2000, p. 547). This mental state was

argued to be similar to the phenomenology illustrated in the oft-cited "butcher on the bus" scenario (Mandler, 1980). A similar analogy to this scenario was offered by Whittlesea and Williams (2000). In their hypothetical situation, a clerk with whom one usually interacts in a store could produce feelings of familiarity if he is not immediately identified as the store clerk when he is encountered in a different context, for example, on a bus. Whittlesea and Williams argued that in this situation, the perceptual processing of the clerk's face would be fluent, but such fluency in processing runs counter to the normative fluency one would expect when processing the face of a stranger. It is this perceived discrepancy between actual and expected fluency which produces a powerful feeling of familiarity, and a sense that remembering is incomplete (Whittlesea & Leboe, 2000). The phenomenology associated with this sense of incomplete remembering was described to be powerful, and it usually occurs when the rememberer is in some degree of doubt (Whittlesea & Williams, 2000). This sense of familiarity may also be accompanied by a prevailing affect, as exemplified in the following verbalisation – "I cannot name the person, or say from where I know [him], but *I'm sure* I've seen [him] before; and I may be right or wrong." (Whittlesea & Williams, 1998, p. 142, own italics). In contrast, when there is no discrepancy between expected and actual fluency, this powerful feeling of familiarity is absent. This is illustrated by a counter-example to the clerk-on-the-bus scenario – one does not have the same pressing feeling of familiarity when one encounters one's spouse in the kitchen at home (Whittlesea & Williams, 2000). In this case, although the perceptual processing of the spouse's face is undoubtedly fluent, this fluency conforms with one's expectations and therefore is unsurprising, and hence no feelings of familiarity would be experienced.

The focus of Whittlesea and Williams's investigations (1998, 2000, 2001a, 2001b) lay in the precise sense of familiarity one would feel when remembering is uncertain and incomplete. In turn, the discrepancy-attribution hypothesis was proposed as a mechanism underlying this particular phenomenology. In studies devised by Whittlesea and Williams, such as those involving pseudohomophones (PHRAUG) and the hension effect (as described above), the aim was to formulate predictions based on the discrepancy-attribution hypothesis in order to recreate or eliminate these feelings of familiarity under experimental contexts. Objectively,

feelings of familiarity are assumed to be reflected by the production of false alarms in recognition tests.

In the next section, some of the supporting evidence provided by Whittlesea and his colleagues for the discrepancy-attribution hypothesis will be presented (e.g., Whittlesea & Williams, 2000, 2001a, 2001b). In later investigations, the discrepancy-attribution hypothesis and the SCAPE framework were expanded by Whittlesea to encompass other memory (e.g., cued-recall – Whittlesea & Leboe, 2003) and decision-making processes (e.g., item classification and generation – Whittlesea & Leboe, 2000; judgments of context reinstatement – Leboe & Whittlesea, 2002). However, because the central interest of the current thesis is on false recognition, the discussion below will be restricted to studies that are pertinent to this topic.

1.10 Other Supporting Evidence for the Discrepancy-Attribution Hypothesis – Perceptions of Discrepancy, Coherence, and Incongruity

In the original conceptualisation of the discrepancy-attribution hypothesis, Whittlesea and Williams (1998) argued that the perception of discrepancy – the antecedent to fluency misattribution and false feelings of familiarity – is produced when there is a mismatch between expectations and actual performance. In the context of the hension effect, Whittlesea and Williams argued that this discrepancy arises from the way that regular nonwords such as HENSION are easy to pronounce, but participants are oblivious to the source of this fluency, and instead regard these regular nonwords as meaningless and therefore should-be-nonfluent items. This view was argued to be substantiated by measures obtained from a rating task, where one group of participants were asked to judge the ease of pronunciation of natural words (CURTAIN), regular nonwords (HENSION) and irregular nonwords (STOFWUS) items, while another group of participants judged whether these items were structurally similar to English words (Whittlesea & Williams, 2001a, Experiment 1). Regular nonwords were judged to be easy to pronounce on 91% of the trials, not significantly different from the figure for natural words (99%). However, regular nonwords were rated significantly lower than natural words in terms of similarity to English words (57% versus 84%). Irregular nonwords were rated low on both of these measures (35% and 14% for pronunciation and similarity respectively). Thus, the magnitude of difference between the pronunciation ease and similarity to English

ratings was the greatest for regular nonwords, and this was taken to be convergent evidence for the existence of discrepancy which is inherent in the processing of regular nonwords.

Another source of convergent evidence sought by Whittlesea and Williams was based on the rationale that if the perception of discrepancy was eliminated, incidence of false recognition would be reduced. To this end, these researchers introduced experimental paradigms which purportedly induced other types of perception during the course of stimulus processing – coherence and incongruity (e.g., Whittlesea, 2002a, 2002b, 2003). The distinction among the three different perceived states (i.e., coherence, incongruity, discrepancy) was illustrated through the following: "Sheila enrolled her son in an academy of music. The teachers quickly identified him as a child progeny." (Whittlesea, 2002b, p. 326). Coherence is said to be experienced if the reader perceives the example as well-formed, in that all aspects appear to be in accord with an overall theme. A sense of coherence would allow the reader to continue with processing the ongoing flow of stimuli in the environment. Other readers might realise that the word "progeny", meaning "offspring", should be more appropriately replaced by "prodigy", meaning "a gifted child". These readers would experience a sense of incongruity, as an error is detected. The flow of processing is consequently disrupted in order for error correction to occur. The final type of perception, that of discrepancy, would be experienced when the reader finds the processing of the sentence to be "surprisingly well" or "surprisingly poor" (Whittlesea, 2002a, p. 98), but cannot pinpoint the cause for this feeling of strangeness. As with incongruity, processing flow is suspended when discrepancy is experienced. Moreover, it was argued that it is this sense of discrepancy which initiates an attributional process whereby the oddness associated with the processing would be ascribed to a source beyond the event itself. For instance, the reader might attribute this sensation of strangeness to mental fatigue (e.g., if the processing seemed "surprisingly poor") or to a prior encounter of the stimulus (e.g., if the processing seemed "surprisingly well", Whittlesea, 2002a, 2002b).

On the basis of these definitions, it was argued that coherent or incongruent processing would not generate false feelings of familiarity (e.g., Whittlesea, 2004). Accordingly, the high level of FA rate observed for regular nonwords in the hension

effect paradigm would be reduced if the processing of these items were made to be coherent or incongruent, rather than discrepant (Whittlesea & Williams, 2000). Two experiments are directly relevant to this proposition. First, to test the effect of coherent processing on the recognition performance of regular nonwords, Whittlesea and Williams (2000, Experiment 4) required participants to perform a rhyme judgment task prior to making the recognition decision. The implementation of the rhyme judgment task was motivated by the following line of reasoning. The cause for discrepancy in the processing of regular nonwords may stem from their lack of meaning. That is, processing ease experienced for regular nonwords violates the expectation that fluently pronounced orthographical units should also be meaningful. Put another way, participants might assume that if an item is fluently pronounced, something more could be done to this item. In the case of natural words for example, meanings associated with the item could be produced. It followed then that if participants were allowed to perform another task on the regular nonwords (other than producing their meaning), the processing of these items would be perceived as coherent, not discrepant (Whittlesea & Williams, 2000).

In Whittlesea and Williams's (2000) rhyme judgment experiment, participants studied a list containing only regular nonwords. In the subsequent recognition test (also containing regular nonwords only), half of test items were presented in isolation, as in the standard hension effect paradigm. Participants were asked to pronounce these items and make a recognition decision. In the other half of the trials, the test item (e.g., PINGLE) was followed immediately by two novel nonwords (e.g., BINGLE, PINGET), of which only one rhymed with the test item. On these trials, participants were required to pronounce the test word, and then select the rhyming nonword before making the old/new recognition judgment for the test item. The FA rate was found to be significantly lower on trials where a rhyming judgment was required (.23), than on trials without the rhyming task requirement (.28). This finding was interpreted as supportive evidence that ordinarily, the processing of regular nonwords is perceived as discrepant, and this discrepancy would engender feelings of familiarity through an attributional process. By making the processing of regular nonwords coherent, discrepancy and thereby feelings of familiarity were reduced.

Apart from coherence, the perception of discrepancy could also be substituted with incongruity through experimental manipulations. In Whittlesea and Williams (2000, Experiment 5), the perception of incongruity was achieved by presenting items in the context of a sentence stem during test. In this experiment, participants studied both natural words and regular nonwords in isolation. At test, each item was presented at the end of a sentence stem (e.g., "The old priest gave the nuns his ..."). The sense of incongruity would be experienced if the test item was a regular nonword (e.g., "The old priest gave the nuns his HENSION"). In the case of natural words, processing could either be predicted (e.g., "The old priest gave the nuns his BLESSING"), or incongruent (e.g., "The old priest gave the nuns his TUNNEL"). Consistent with the idea that incongruity does not produce feelings of familiarity, FA rates for regular nonwords (.08) and natural words in incongruent contexts (.10) were significantly lower than that for natural words in predictive contexts (.21).

1.11 Varieties of Discrepancy

At first glance, the inflated FA rate obtained for natural words in the context of predictive sentence stems appears to contradict the notion that feelings of familiarity are caused by discrepancy in item processing. If anything, the completion of a sentence stem with a sensible natural word creates a sense of coherence, not discrepancy, and hence false feelings of familiarity should not be induced. To counter this argument, Whittlesea and Williams (2001b) maintained that discrepancy does exist for natural words in the context of predictive sentence stems, and further suggested that this sense of discrepancy hinges on the presence of a pause between the sentence stem and the terminal natural word item. Rather serendipitously, these researchers first speculated the importance of the pause in a previous experimental procedure (Whittlesea, 1993), where they noted that the length of presentation of the sentence stem (at a fixed duration of 2 seconds) was usually long enough to allow for a pause after the stem had been read and before the terminal item was revealed. Consistent with this speculation, Whittlesea and Williams showed that false recognition of the terminal natural word was reduced when there was no pause following the predictive sentence stem (2001b, Experiment 1). They argued that this pause crucially allowed participants to form a general expectation about the termination of the sentence, but this expectation was also accompanied by a sense of

uncertainty as while the sentence stem was predictive (e.g., "She cleaned the kitchen floor with a ..."), there were still a number of possible words (e.g., BROOM, MOP, RAG, BRUSH) which could complete the sentence in a meaningful way. When the terminal word (in this case, BROOM) was finally presented, it was suggested that mentally, the participant would shift rapidly to a state of resolution and understanding. Metaphorically speaking, these investigators likened the effect to "waiting for the other shoe to drop" (Whittlesea & Williams, 2001b, p. 17). That is, although the terminal word was anticipated, its revelation was ultimately startling because one did not know for certain when it would occur. With this sudden resolution, the participant would experience surprise, and the perception of discrepancy would follow when this sense of surprise could not be attributed to the current situation. Perhaps in support of this proposition, Whittlesea and Williams showed that only predictive sentence stems (e.g., "She cleaned the kitchen floor with a... BROOM"), but not completely predictive (e.g., "The rolling stone gathers no... MOSS", Experiment 2) or merely consistent (e.g., "She couldn't find a place to put the... BROOM", Experiment 4) sentence stems would produce illusory feelings of familiarity.

In view of the way that discrepancy could be produced in subtly different ways, Whittlesea and Williams (2001b) further argued that there are in fact several "varieties" of discrepancy. For example, "surprising inconsistency" is associated with regular nonwords such as HENSION, where the pronunciation for this item is experienced to be fluent, but there is a failure to produce a corresponding meaning to the item, hence creating a sense of inconsistency. "Surprising redintegration" is experienced in pseudohomophones such as PHRAUG – participants are assumed to expect a lack of meaning for these nonwords, yet upon the pronunciation of these nonwords, participants would realise that the phonology of these items correspond to real words with real meanings. "Surprising coherence" refers to the way the item in question fits surprisingly well in a given context[4]. In "surprising coherence", the source of discrepancy comes from the way an expectation formed in a given situation

[4] Relevant to the concept of "surprising coherence" is *integrality* (Whittlesea, 2004) – introduced as another perception which can be experienced during item processing. Like discrepancy, integrality was argued to produce feelings of remembering, but it was implicated specifically in the context of memory for whole sentences, rather than individual words in isolation. Because of this, Whittlesea's work relating to integrality will not be discussed in detail here.

is *affirmed*, rather than *violated*. Anticipating the confusion between the notions of "coherence" and "surprising coherence" (which was held to be a variety of discrepancy), Whittlesea and Williams likened these two perceptions to those that would be experienced by a teacher receiving outstanding work from an excellent versus a mediocre student. In the former case, the work would be seen as "good", but in the latter case, it would be "surprisingly good". These authors further argued that "surprising coherence" was the type of discrepancy accountable for the way a natural word seemed to fall fittingly at the end of a predictive sentence stem, and consequently the high FA rate associated with these items (Whittlesea & Williams, 2000).

1.12 Problems with the Discrepancy-Attribution Hypothesis

The complexity in the definition of discrepancy highlights the principal problem faced by the discrepancy-attribution hypothesis. As conceded by the theorists themselves, the hypothesis is "somewhat vague" in its structure and predictions (Whittlesea & Williams, 2000, p. 559). Part of the vagueness might stem from the way that perceptions of coherence, incongruity, and the various kinds of discrepancy are not directly measurable. Indeed, these perceptions are assumed to be unconscious and could only be inferred from behavioural data (Whittlesea & Williams, 2000). This assumption poses a problem of circularity when investigators attempt to evaluate the hypothesis: Is the context discrepant, or is it judged to be discrepant just because more false alarms had been produced in this condition? The same difficulty is also apparent in deciding whether processing was coherent or "surprisingly coherent", as it seems that this dilemma can only be confidently resolved by making inferences from FA rates – a high FA rate would indicate that surprising coherence, rather than plain coherence, was experienced. It could be seen that in the search for theory-conforming evidence, investigators could run into the danger of re-interpreting experimental paradigms and accounting for their findings in a post hoc fashion. Consequently, the hypothesis could fall into the trap of being ultimately unfalsifiable.

Another potential criticism directed towards the discrepancy-attribution hypothesis is that many of the findings which evidently supports the hypothesis could be explained through more parsimonious accounts, without the necessity of proposing a multitude of other concepts (i.e., varieties of discrepancy, coherence, incongruity,

etc.) to explicate the findings. For instance, in the rhyme judgment experiment detailed above (Whittlesea & Williams, 2000, Experiment 4), the decrease in false recognition for regular nonwords on trials requiring rhyme judgments could be due to the possibility that whatever fluency experienced for these items was attributed to the preceding rhyme judgment task, rather than a source in the past. A finding from Jacoby and Whitehouse (1989) is particularly relevant here. As mentioned earlier, these researchers showed that "old" judgments for items would increase if they were preceded by a matching prime, rather than a nonmatching or control prime. Importantly, it was also found that the presentation duration of the matching prime was critical in producing this effect – increase in "old" responses in the matching-prime condition was only observed if the prime duration was short enough (16 ms) for participants to remain "unaware" of the prime's existence. When the prime duration was sufficiently long enough (600 ms) for participants to be "aware" of the prime's presence, the probability of "old" judgments was actually *lower* in the matching-prime than in the nonmatching- or control-prime condition. Jacoby and Whitehouse reasoned that for both unaware and aware participants, the matching prime enhanced the processing fluency of the subsequent test item. However, for the unaware group, this fluency was attributed to pastness, thus creating feelings of familiarity. In contrast, the aware participants attributed this same fluency to a source deemed to be most plausible, namely the preceding matching prime itself, as it was clearly shown for 600 ms. In a similar way, in Whittlesea and Williams's rhyme judgment paradigm, one of the alternatives (e.g., BINGLE) in the rhyme judgment task always resembled the regular nonword test item (e.g., PINGLE) in both phonology and orthography. It is therefore reasonable to suggest that the fluency experienced for the test item was likely to be attributed to the rhyming nonword, rather than to the past. Consequently, the FA rate would be lower in trials with a preceding rhyme judgment task than those without. Thus, the notion of coherence need not be implicated in explaining Whittlesea and Williams's data.

Likewise, results obtained for regular nonwords in the sentence stem paradigm (Whittlesea & Williams, 2000, Experiment 5) could be explained without invoking the notion of incongruity. The low FA rate for regular nonwords, presented in the context of meaningful sentence stems, might be caused purely because the processing of these items was simply not perceived to be fluent, and as such, there was no

fluency available to be attributed to the past. This argument could be supported by the fact that regular nonwords were pronounced significantly more slowly than natural words. Because the sentence stems used were predictive in that a meaningful word was expected to complete the sentence, a terminal regular nonword would therefore be perceived as particularly nonfluent.

The lack of fluency experienced for regular nonwords in the sentence stem paradigm may also be partly due to the absence of irregular nonwords in the experiment. Indeed, Whittlesea and Williams (e.g., 2000, 2001a) had generally excluded irregular nonwords in the follow-up experiments centring on the hension effect paradigm, preferring to concentrate on the FA-rate difference between regular nonwords and natural words. It is conceivable that in the original hension effect experiment (Whittlesea & Williams, 1998), regular nonwords were experienced to be fluent because comparatively, irregular nonwords were remarkably low in fluency. In contrast, in an experiment where the only other items available for comparison were natural words, regular nonwords would likely be perceived as nonfluent. This perceived nonfluency of regular nonwords would therefore undermine Whittlesea and Williams's (2000) argument that false recognition for these items could be reduced by rendering their processing incongruent.

Another consequence of removing irregular nonwords from the experimental design was that the definition of the hension effect itself became increasingly unclear in Whittlesea and Williams's (2000, 2001a) later investigations on the effect. In the original report of the hension effect, the emphasis was placed not only on the FA-rate difference between regular nonwords and natural words, but also on the difference between regular and irregular nonwords. Thus, it was puzzling why Whittlesea and Williams had decided to abandon irregular nonword items in their follow-up studies. Indeed, in these later experiments conducted by Whittlesea and Williams (e.g., 2000, Experiment 3), the hension effect appeared to be defined simply as instances of false recognition produced for regular nonwords, and hence the effect was claimed to be observed ("the effect was as large", p. 559) even when the study and test list consisted entirely of these items, where no other item group was provided as a benchmark for comparison.

1.13 Concluding Remarks

In this introductory chapter, a brief review of the research on fluency attributions was presented. In relation to recognition memory, the fluency experienced during the course of stimulus processing was argued to be a basis by which feelings of remembering for that stimulus could be produced (e.g., Jacoby et al., 1989). Moreover, according to this attributional account, false recognition would occur if fluency in current processing was caused from some factor other than prior experience, but was misattributed incorrectly to a source in the past. However, it was noted that the relationship between fluency and the phenomenology of remembering is less straightforward than first thought. Experience of fluency does not inevitably translate to feelings of remembering, and subsequently judgments of old. To account for the indirect relationship between fluency and claims of remembrance, the dual-process model of recognition memory (e.g., Jacoby & Dallas, 1981; Mandler, 1979, 1980) was proposed where it was suggested that recognition could be performed on the basis of either recollection or familiarity. Further, the assumption of the dual-process model is that familiarity is the primary route by which false recognition is produced (e.g., Jacoby, 1991).

As an alternative to the dual-process model, the discrepancy-attribution hypothesis was offered by Whittlesea and Williams (1998) to explain false recognition, and its indirect relationship with processing fluency. The hypothesis has its basis in the SCAPE framework (Whittlesea, 1997, 2003), where memories are assumed to be reconstructed through two stages: production – where a cognitive operation is performed on an encountered stimulus, and evaluation – where the cognitive operation is assessed and the phenomenology of remembering is generated. Whittlesea and Williams argued that false feelings of familiarity – a phenomenology analogous to incomplete and uncertain remembering – originate from a perception of discrepancy during the processing of the stimulus. As support for their argument, these researchers demonstrated the hension effect, where the rate of false recognition was found to be higher for regular nonwords than both natural words and irregular nonwords. It was reasoned that discrepancy was experienced during the processing of regular nonwords, because these items could be pronounced fluently, but were meaningless. This violation of expectation (that fluently pronounced stimuli should be

meaningful) associated with the processing of regular nonwords would in turn prompt participants to attribute fluency to pastness, thus creating (false) feelings of familiarity.

In relation to research on recognition memory, Whittlesea and Williams's (1998) discrepancy-attribution hypothesis provides an interesting and plausible conjecture on the cause of false recognition. In subsequent work by these authors (2000, 2001a, 2001b), and by Whittlesea himself (2002a, 2002b, 2004), the theory was extended with the proposal of other perceptions associated with stimulus processing (viz., coherence, incongruity, integrality), and other "varieties" of discrepancy. The distinctions among these concepts are, in some cases, exceedingly subtle but are critical in the sense that they predict different experimental outcomes. On the one hand, the expansiveness of these investigations signifies the far-reaching potential of the discrepancy-attribution hypothesis and the SCAPE framework in providing a valuable account of many of our everyday phenomenological experiences. On the other hand, the proliferation of new concepts may attract criticisms that the hypothesis is marked by vagueness, is lacking in parsimony, and is ultimately an unfalsifiable, invalid scientific theory (Popper, 1959).

In view of these criticisms, and given that the hension effect constitutes a founding piece of evidence for the discrepancy-attribution hypothesis, it may be worthwhile to conduct a careful re-examination of the efficacy of using the hypothesis to explain the FA rate pattern shown in the hension effect. This is the first objective of the current thesis. An ensuing objective is to explore other possible factors involved in producing the hension effect. As will be investigated in later chapters of the thesis, one mechanism – that of memorability-based rejection of new items – may exert significant influence on the levels of FA rates obtained in the hension effect, and as such, this mechanism should form an essential part in any model pertaining to recognition memory.

Chapter 2

Of the vast number of experiments conducted by Whittlesea and his colleagues, the hension effect (Whittlesea & Williams, 1998) was arguably one of the most robust, and hence one of the most cited findings. Indeed, within two years after the initial report of the phenomenon, the effect was claimed to have been replicated on 14 other occasions (Whittlesea & Williams, 2000).

As detailed in the previous chapter, the hension effect became the impetus in the formulation of the discrepancy-attribution hypothesis (Whittlesea & Williams, 1998). Subsequent investigations by Whittlesea and Williams (2000, 2001a, 2001b) were marked by an increase in both the complexity, and perhaps also the vagueness of the hypothesis. In its most basic form, however, the discrepancy-attribution hypothesis provides a cogent explanation for the hension effect. It was argued that the processing of the regular nonwords is initially fluent (as reflected by the short pronunciation latencies for these items), this initial fluency in turn creates an expectation that more processing could be performed with the item, for example, that its meaning could be retrieved. Because the item was in fact a nonword, no meaning could be generated. This results in a sense of "surprising failure" because the initial expectation and the following actual outcome are discrepant[5]. It is the perception of discrepancy which triggers an unconscious attributional process, whereby the source of fluency experienced in the initial processing is attributed to the source most plausible to the participant – in this instance, the past – and thus a conscious feeling of familiarity is generated. Such discrepancy does not exist for natural words (which are fluent and meaningful) or irregular nonwords (which are nonfluent and meaningless). This consistency between expectation (which is formed on the basis of processing fluency) and subsequent outcome (which is dependent on the item's meaningfulness)

[5] In the PHRAUG experiment (Whittlesea &Williams, 1998, Experiment 1, see section 1.8 for details), the resultant feeling is "surprising success". The initial processing of the pseudohomophone is nonfluent, leading to an expectation that nothing more could be performed on the item. Therefore, the ensuing realisation that the item's pronunciation is associated with a real word's meaning constitutes "surprising success". Thus, it is not of importance whether it is success or failure that eventuates, the critical element is the surprise of the experience, that is, the initial expectation and the actual performance are discrepant.

does not create a sense of discrepancy, and hence no attribution of fluency and no feelings of familiarity would result.

2.1 Testing the Discrepancy-Attribution Hypothesis: Manipulations on Fluency Evaluation, Processing Fluency, and Meaningfulness

The following two chapters will describe four experiments which were conducted to evaluate the applicability of the discrepancy-attribution hypothesis for the hension effect. These experiments were designed on the rationale that the hension effect would be eliminated along with the removal of the perception of discrepancy associated with regular nonword items. Hypothetically, the sense of discrepancy experienced for these items could be abolished through manipulations at three different stages during item processing: (a) the evaluation of processing fluency for the item, (b) the perception of actual fluency achieved for the item, and (c) the retrieval of the item's meaning. In each of the four experiments, a manipulation would therefore be imposed on one of these stages, with the intention to eliminate discrepancy, and hence the hension effect, which was defined as the elevated level of FA rates for regular nonwords, relative to natural words and irregular nonwords.

In Experiments 1 and 2, a feedback manipulation was introduced which targeted the evaluative component of item processing. According to the discrepancy-attribution hypothesis and the SCAPE framework, once an item had been initially processed, participants would then evaluate the fluency experienced during processing and compare this evaluation with the expectations formed for the item. The feedback manipulation then, was imposed with the purpose of influencing the evaluation of processing fluency. Specifically, in some conditions, false feedback would be given, in an attempt to bias participants into perceiving their processing of regular nonwords as nonfluent. Thus, under this perception of nonfluency, the subsequent failure in producing meanings for regular nonwords would no longer be discrepant, and hence an increased rate of false recognition would not be observed for these items.

In Experiment 3, the manipulation would directly focus on the actual level of processing fluency produced for the items in the hension effect. The explication of the effect, as provided by the discrepancy-attribution hypothesis, is grounded on the assumption that the processing of regular nonwords is initially fluent. In Whittlesea

and Williams (1998), this fluency in processing was argued to be evinced by the way that pronunciation latency was fast for regular nonwords, particularly in comparison to irregular nonwords. In Experiment 3, a preceding lexical decision task (LDT) was included in the test procedure in order to disrupt the processing fluency normally generated for regular nonwords. The rationale was that if the processing of regular nonwords itself was made nonfluent, there would be no perception of discrepancy because nonfluency is connected with these items' meaninglessness.

While the focus of the manipulations devised in Experiments 1 – 3 was directed at processing fluency, Experiment 4 (which will be reported in Chapter 3) employed a manipulation which centred on the meaninglessness of regular nonwords. According to the discrepancy-attribution hypothesis, processing fluency experienced for regular nonwords creates the expectation that these items would be meaningful. The eventual failure to retrieve meanings for these items is therefore discrepant to this expectation. Thus, if regular nonwords are assigned individual meanings, discrepancy and the resultant feelings of familiarity would be eliminated. This prediction was tested in Experiment 4.

2.2 Experiment 1: Manipulating the Evaluation of Fluency (Feedback on Latency)

The SCAPE framework, and in turn the discrepancy-attribution hypothesis, places an emphasis on the role of performance evaluation in the generation of feelings of familiarity. In the hension effect, the processing fluency of regular nonwords was generally evaluated to be fluent, and therefore discrepant with the reality that these items were meaningless. It follows then that if the processing of these regular nonwords were evaluated to be *nonfluent*, then discrepancy would be eliminated. Consequently, in the absence of discrepancy, no unconscious attributional process would take place to generate false feelings of familiarity.

Regular nonwords could be contrasted with irregular nonwords in that the processing fluency for irregular nonwords is normally regarded to be nonfluent. Because these items are also meaningless, there would be no perception of discrepancy, and this is reflected by the low levels of FA rates for these items. Hypothetically then, if the processing of irregular nonwords is evaluated to be *fluent*,

then a sense of discrepancy would follow, precipitating in a misattribution of fluency to pastness.

The above two propositions formulated for regular and irregular nonwords were examined in Experiment 1, where the fluency evaluation for these items would be modified such that processing for regular nonwords would be considered as nonfluent, and for irregular nonwords, fluent. The procedure of the test phase would be largely identical to that in the original paradigm (Whittlesea & Williams, 1998) – on each test trial, participants were required to first pronounce the test item, and then make a recognition judgment on the item. The exception in the present experiment is that on each test trial, participants would be given feedback relating to the time they had taken to read aloud the test item. This "pronunciation duration" measure would therefore constitute an index to the item's processing fluency. It can seen that the veracity of the feedback could be manipulated such that processing could be perceived as either more, or less fluent than in actuality. Thus, false feedback could be given to regular nonwords to render them nonfluent (i.e., as would be reflected by a longer pronunciation duration), and to irregular nonwords to render them fluent (i.e., as would be reflected by a shorter pronunciation duration).

In Whittlesea and Williams's hension effect experiments (1998, Experiments 3, 5 and 6), pronunciation latency (time taken to initiate pronunciation) was provided as an objective measure for processing fluency. Averaged across those experiments, the data showed that the mean pronunciation latency for regular nonwords was approximately 350 ms faster than that for irregular nonwords. This result from Whittlesea and Williams was used here to determine the adjustment to be made in the false feedback condition. For regular nonwords, the false feedback would indicate a pronunciation duration that was 350 ms *slower* than in actuality, and for irregular

nonwords, 350 ms *faster* than in actuality[6]. Thus, for example, if the participant took 1,000 ms to pronounce a regular nonword, the false feedback would show that the pronunciation duration was 1,350 ms. In contrast, if the participant in reality took 1,350ms to pronounce an irregular nonword, the false feedback would show that the pronunciation duration was 1,000 ms. Because Whittlesea and Williams found that processing was on average, 350 ms faster for regular than irregular nonwords, the false feedback in effect would reverse this pattern such that irregular nonwords would seemingly be processed, on average, 350 ms more fluently than regular nonwords.

Two additional feedback conditions – "true feedback" and "no feedback" – were incorporated into the design of Experiment 1 to serve as points of comparison for the false-feedback condition. Thus, both regular and irregular nonword test items were presented in all three, true-, no-, and false- feedback conditions. Because the false-feedback condition did not apply to natural words, these items were presented in the true- and no-feedback conditions only. On a true-feedback test trial, the true pronunciation duration attained by the participant for the test item would be given as feedback. On a no-feedback test trial, no information was given regarding the item's pronunciation duration, and hence the procedure here would be most similar to that employed by Whittlesea and Williams (1998). In this way, the no-feedback condition would serve to replicate the hension effect.

It was therefore predicted that in the no-feedback condition, regular nonwords would produce a higher FA rate than both natural words and irregular nonwords. The same FA rate pattern was expected in the true-feedback condition because the feedback here should not eliminate the perception of discrepancy for regular nonwords. In the false-feedback condition, however, the removal of discrepancy for regular nonwords was expected to be accompanied by a reduction of FA rate for these

[6] Unlike Whittlesea and Williams (1998), pronunciation duration, rather than latency, was used as a measure for processing fluency in the current experiment. In Whittlesea and Williams's experiments, pronunciation latency data were collected by requiring participants to press a key as they were initiating pronunciation. In the current experiment, it was reasoned that this additional key-press requirement might distract participants from giving full attention to the feedback. Therefore, it was decided that measures of fluency would be recorded by the experimenter. Consequently, total duration, rather than onset latency was preferred because it was anticipated that the experimenter would be more accurate in recording the endpoint, rather than the onset of the participant's pronunciation.

items (relative to the FA rates produced in the true- and no-feedback condition). In contrast, false feedback for irregular nonwords was argued to invoke a sense of discrepancy for these items. Thus, it was hypothesised that FA rates for irregular nonwords would be higher in the false-feedback, than in the true- and the no-feedback condition. Together, these predictions made for regular and irregular nonwords would culminate in the elimination of the hension effect in the false-feedback condition, with the FA rate of irregular nonwords being equivalent to, or perhaps even higher than, that for regular nonwords.

Whether these predicted outcomes would eventuate might depend on an essential factor – that participants would incorporate feedback information when making their recognition decisions. In two previous experiments carried out by Whittlesea and Williams (1998, Experiments 5 and 6), externally presented information was also employed in an attempt to "debias" participants such that they would be less inclined, in the case of regular nonwords, to misattribute fluency to the past. In these experiments, participants were informed that the regular nonwords were constructed specially to resemble English words, and would therefore be easily processed. They were further advised to discount the ease of processing for these items, lest their recognition judgments would be unduly biased by this factor. Even when aided by test labels, which indicated that a regular nonword was an "easy nonword" (natural words and irregular nonwords were presented with the labels "word" and "hard nonword" respectively), the hension effect was nonetheless obtained. It was argued that the perception of discrepancy, and the subsequent feelings of familiarity experienced for regular nonwords are "cognitively impenetrable" (Whittlesea & Williams, 1998, p. 155). Analogically speaking, Whittlesea (2002) suggested that "just as being told that a candy bar contains empty calories does not stop it from tasting good, so being warned that hension is designed to be easy to say does not stop it from feeling familiar" (p. 104). Thus, the subjective assessment of fluency for regular nonwords might not be easily swayed by an objective indicator of fluency. For the feedback manipulation to exert its effects, participants must therefore demonstrate an ability to apply the externally presented indicator of fluency in a strategic manner. If fluency misattribution is an automatic and unconscious process that is immune to external influences, the hension effect would be observed regardless of the type of feedback given.

2.2.1 Method

Participants. Twenty-seven psychology undergraduates from the University of Southampton participated in this experiment in return for course credits. For all participants, English was the only language they could speak fluently.

Materials and Design. The materials used in Experiment 1 consisted of the 60 items from three categories, as devised by Whittlesea and Williams (2000). These are listed in Appendix A. The three categories are natural words (e.g., CURTAIN, DAISY), regular nonwords (e.g., HENSION, BARDEN), and irregular nonwords (e.g., STOFWUS, LERTISP). Additionally, six items (two for each category, see Appendix A) were constructed here to serve as practice items.

For each participant, half (i.e., 30 items) from each category were randomly selected to be presented during the study phase. Half (i.e., three) of the constructed practice items were also presented at the start of the study phase. Thus, apart from the practice items, there were 90 study items in total. At test, all six practice items were presented at the beginning. These were followed by all 60 items from each category, resulting in a total of 180 test items. For the regular and irregular nonword categories, a third (i.e., 20) were randomly assigned to each of the feedback conditions – true feedback, no feedback and false feedback. For natural words, two thirds of the items (i.e., 40) were randomly assigned to the true-feedback, and the remaining third (i.e., 20) to the no-feedback condition. In each of the feedback conditions, it was ensured that within each item category, exactly half of the items were old, and half were new. Item assignment to old versus new status, and to feedback conditions, was freshly performed for each participant. For the 6 practice test items, 4 were given true feedback, and 2 were given no feedback.

Procedure. Participants were tested individually in a quiet cubicle. The experiment was programmed using the software Revolution 2.0.2 and was presented to participants via a computer monitor. Instructions given before the study phase were presented on the computer screen and reiterated verbally by the experimenter to prevent misunderstanding. In these instructions, participants were informed that the study list would include both English and non-English items. They were also asked to remember each study item for a later recognition memory test, and thus the learning

of the study items would be intentional (see Appendix B for the standard set of study phase instructions). Following these instructions, the study phase commenced with the three practice items presented in a fixed order across participants, who were not informed of the practice status of these items. The 90 study items followed immediately after the practice items, and these were presented in a freshly randomised order for each participant. All items in the study phase (including the practice items) were presented for 1 s each, with an inter-trial interval (ITI) of 1 s. Additionally, all study items were presented in Arial font, in lowercase, in black and on a white background at the centre of the computer screen. Each letter measured approximately 0.8 cm in width and 1.2 cm in height.

Immediately after the end of the study phase, the instructions for the test phase were presented. These instructions, like those for the study phase, were both shown visually on the computer screen and given verbally by the experimenter. Participants were informed that the test would consist of both English and non-English words – some of which were presented previously in the study phase, whereas others were not. Participants were further instructed that for each item in the recognition test, their first task was to pronounce the item as quickly and accurately as possible. They were also told that for some trials, feedback would be given regarding the time they had taken to pronounce the item, whereas for other trials, they would not receive this feedback[7]. Following this, participants were told that they would see the same item again, but this time they were to respond "old" to items that were previously studied and "new" to items that were not previously studied. This response was to be given verbally to the experimenter.

The test phase followed immediately after the test instructions with the presentation of six trials containing the practice items. These trials were given in a fixed order across participants, who were not informed of the practice status of these

[7] For the last nine participants in this experiment, there was an additional requirement to read the feedback aloud to the experimenter. This procedural change was imposed following the concern that without this requirement, the feedback was not attended to by the participants. Subsequent analyses showed that the recognition performance obtained from these nine participants did not differ from that produced by the first 18 participants who did not read out the feedback. Thus, data from all 27 participants were pooled together and analysed as a single group (see details in the Results section, 2.2.2).

trials. During the test phase, each trial began with the presentation of the test item, which was displayed in the same manner (in terms of location on the screen, size, colour, etc.) as items in the study phase. A prompt "Read Word Aloud" was also presented towards top of the screen to remind participants of their task. The time taken by participants to pronounce the item was recorded by the experimenter via button press on a Cedrus RB-620 response pad, which in turn was connected to the serial port of the computer. Once the pronunciation duration was recorded, the item was replaced on the screen by the feedback. On a true-feedback trial, the time taken by the participant to pronounce the item (in milliseconds) was presented in red at the centre of the screen, along with an accompanying label "Time Taken" displayed above the feedback. In the false-feedback condition, 350 ms were added to the true pronunciation duration of a regular nonword, and 350 ms were subtracted from the true pronunciation duration of an irregular nonword, to create a seemingly slower and a seemingly faster pronunciation duration respectively. The false feedback was displayed in the same manner as the true feedback. The on-screen duration for both true and false feedback was 2 s, and for trials in the no-feedback condition, a blank was presented to the participant for the same time interval. Following this, the test item reappeared, along with the prompt "Old or New?" to remind participants that they were now required to make a recognition judgment for the item. The response was made verbally by the participant, and once again this was recorded by the experimenter via the response pad. On each test trial, the item remained on the screen until the participant's old/new response had been recorded, and an ITI of 1 s was deployed before the next test trial began. For each participant, the test items were presented in a uniquely randomised order, but with the additional constraint that no more than 3 consecutive trials would be associated with the no-feedback condition.

Due to the length of the recognition test, participants were offered an opportunity to take a short one minute break at the half-way mark of the test (after 90 test trials). Throughout the course of the test phase, a counter was also displayed at the bottom right-hand corner of the screen to inform participants of the number of test trials remaining. In its entirety, the experiment lasted approximately 30 minutes. Upon the completion of the session, participants were thanked, debriefed and dismissed.

2.2.2 Results

Statistical analyses were performed on three sets of dependent measures obtained from Experiment 1: (a) actual pronunciation durations obtained for the items, (b) hit rates, and (c) FA rates.

Pronunciation Duration. Table 1 shows the mean pronunciation durations for each item category, arranged in terms of item status (old or new). A 3 (item: natural word/ regular nonword/ irregular nonword) x 2 (status: old/ new) repeated-measures ANOVA (Analysis of Variance) revealed a significant main effect of item, $F(2, 52) = 121.19$, $p < .001$, $MSE = 70816.50$, $\eta^2 = .823$. Post-hoc comparisons with Bonferroni adjustment to the alpha ($\alpha = .0167$) showed that the item main effect was due to significantly shorter pronunciation durations produced for natural words ($M = 1117$) than for regular nonwords ($M = 1219$ ms), $t(26) = 7.30$, $p < .001$, $SE = 13.95$, $\eta^2 = .758$. In turn, mean pronunciation duration for regular nonwords was faster than that for irregular nonwords ($M = 1853$), $t(26) = 11.10$, $p < .001$, $SE = 57.13$, $\eta^2 = .879$. The main effect of item status was found to be approaching significance, $F(1, 26) = 3.24$, $p = .083$, $MSE = 2447.30$, $\eta^2 = .083$. The mean pronunciation duration was longer for new items ($M = 1403$) than for old items ($M = 1389$). The interaction was non-significant, $F(2, 52) = 2.12$, $p > .10$.

Table 1. Experiment 1 ($N = 27$): Pronunciation durations (in ms) for each item category, arranged in terms of item status (old/new). Standard deviations are in parentheses.

	Old		New	
Natural	1118	(139)	1117	(126)
Regular	1215	(171)	1223	(178)
Irregular	1836	(434)	1870	(418)

Hit and FA rates from Experiment 1 are presented in Tables 2 and 3 respectively. As these tables show, because the false feedback condition was not

applied to natural words, the cell corresponding to this condition is empty. Consequently, Experiment 1 could be thought of as comprising two separate designs. First, the exclusion of natural word data would result in a 2 (item: regular/ irregular nonwords) x 3 (feedback: true/ no/ false) repeated-measures design. Second, if data from the false feedback condition are excluded, the outcome is a 3 (item type: natural word/ regular nonword/ irregular nonword) x 2 (feedback: true/ no) repeated-measures design. These two designs were used as the basis for the analyses performed on hit and FA rates.

Table 2. Experiment 1: Means (and standard deviations) of hit rates for the three item types, under true-, no-, and false- feedback conditions ($N = 27$).

	True	No	False
Natural	.72 (.21)	.72 (.18)	–
Regular	.66 (.22)	.66 (.20)	.70 (.18)
Irregular	.52 (.21)	.50 (.20)	.56 (.23)

Hit Rate. A 2 (item: regular nonword/ irregular nonword) x 3 (feedback: true/ no/ false feedback) repeated-measures ANOVA conducted on hit rates showed a significant main effect of item, $F(1, 26) = 22.27$, $p < .001$, $MSE = .040$, $\eta^2 = .461$, reflecting the fact that the average hit rate was higher for regular nonwords ($M = .67$) than for irregular nonwords ($M = .52$). Neither the feedback main effect nor the interaction was found to be significant, $F(2, 52) = 1.14$, $p > .30$, and $F < 1$ respectively. Similarly, the 3 (item: natural word/ regular nonword/ irregular nonword) x 2 (feedback: true/ no feedback) repeated-measures ANOVA showed a significant item main effect, $F(2, 52) = 14.08$, $p < .001$, $MSE = .045$, $\eta^2 = .351$, but no other significant main effect of feedback or interaction, both Fs < 1. Post-hoc contrasts ($\alpha = .0167$) revealed that the item main effect in this analysis was driven by a greater average hit rate produced for natural words ($M = .72$) than irregular nonwords ($M = .51$), $t(26) = 4.26$, $p < .001$, $SE = .050$, $\eta^2 = .516$; and for regular nonwords ($M = .66$) than for irregular nonwords, $t(26) = 4.46$, $p < .001$, $SE = .034$, η^2

= .540. The difference in hit rates between natural words and regular nonwords was not significant, $t(26) = 1.56, p > .10$.

Table 3. Experiment 1: Mean FA rates obtained for each item type in the three feedback conditions ($N = 27$). Standard deviations are in parentheses.

	True	No	False
Natural	.19 (.13)	.17 (.12)	–
Regular	.37 (.26)	.37 (.18)	.37 (.21)
Irregular	.27 (.17)	.18 (.14)	.25 (.19)

FA Rate. The first analysis on FA rates – a 2 (item: regular nonword/ irregular nonword) x 3 (feedback: true/ no/ false feedback) repeated-measures ANOVA – showed a significant item main effect, $F(1, 26) = 27.62, p < .001, MSE = .027, \eta^2 = .515$. The FA rate for regular nonwords ($M = .37$) was significantly higher than that for irregular nonwords ($M = .24$). The feedback main effect and interaction were not found to be significant, $F(2, 52) = 1.65, p > .20$ and $F(2, 52) = 1.43, p > .25$ respectively. The second analysis on FA rates – a 3 (item: natural word/ regular nonword/ irregular nonword) x 2 (feedback: true/ no) repeated-measures ANOVA – also produced a significant main effect of item, $F(2, 52) = 18.73, p < .001, MSE = .028, \eta^2 = .419$. In this analysis, however, the main effect of feedback was marginally significant, $F(1, 26) = 3.43, p < .08, MSE = .015, \eta^2 = .116$, because the FA rate was higher under the true-feedback ($M = .28$) than the no-feedback ($M = .24$) condition. The item main effect arose because regular nonwords ($M = .37$) produced a significantly higher FA rate than both natural words ($M = .19$) and irregular nonwords ($M = .23$), $t(26) = 4.97, p < .001, SE = .038, \eta^2 = .592$, and $t(26) = 4.78, p < .001, SE = .030, \eta^2 = .573$ for the respective post-hoc paired-samples t tests ($\alpha = .0167$). Natural words and irregular nonwords did not differ significantly in FA rates, $t(26) = 1.56, p > .10$. The item x feedback interaction was not significant, $F(2, 52) = 2.24, p > .10$.

2.2.3 Discussion

Despite using a different index of processing fluency (i.e., pronunciation duration), the present experiment produced data that are comparable to the pronunciation latency measures obtained by Whittlesea and Williams (1998). In general, processing was more fluent for natural words than regular nonwords, and more fluent for regular nonwords than irregular nonwords. The effect of old/new status, however, was marginal. Nonetheless, the direction of the effect was consistent with the concept of repetition priming (e.g., Jacoby & Dallas, 1981), which predicts that relative to a novel stimulus (i.e., new item), a previously encountered stimulus (i.e., old item) would be processed more fluently when encountered for a second time.

There was, however, a difference between the pronunciation duration data obtained in Experiment 1, and the pronunciation latency data provided by Whittlesea and Williams (1998). Specifically, the regular nonword-irregular nonword difference in pronunciation duration (634 ms) was nearly twice as large as the corresponding difference in pronunciation latency (approximately 350 ms, Whittlesea & Williams, 1998). This finding suggests that the fluency advantage enjoyed by regular nonwords over irregular nonwords may increase in magnitude during the course of item processing. This finding was unanticipated, and somewhat muted the potency of the false feedback condition. Based on Whittlesea and Williams's pronunciation latency data, the 350 ms false feedback modification would reverse the perception of fluency for regular and irregular nonwords such that the latter would be considered as more fluent than the former. Because the difference in pronunciation duration was twice as large as expected, the 350 ms modification in fact rendered the perception of fluency for these two items types to be approximately equal. Nevertheless, the prediction for the FA rate pattern under the false-feedback condition remained unchanged. Under this condition, the feedback indicated that the processing was less fluent than normal for regular nonwords, and more fluent than normal for irregular nonwords. The sense of discrepancy would diminish in the former case, and would arise in the latter case. Consequently, compared to the true- and no-feedback conditions, the FA rate was predicted to be lower for regular nowords, and higher for irregular nonwords in the false-feedback condition.

The data from Experiment 1 did not support this prediction. Indeed, the hension effect was observed across all feedback conditions. Feedback did not appear to have an influence on how processing fluency was evaluated by participants. For regular nonwords, in particular, the FA rate was identical across the three feedback conditions. However, when all three item types were analysed together, there was a main effect of feedback, indicating that the FA rate was elevated in the true-feedback, relative to the no-feedback condition. This finding suggests that feedback information may be strategically incorporated into the recognition process to some extent, although it remained unclear why the presence of true feedback would necessarily increase instances of false alarms.

Apart from the processing fluency data (i.e., pronunciation duration), the hit rate obtained in Experiment 1 also deviated from the pattern typically found in the hension paradigm (e.g., Whittlesea & Williams, 1998, 2000). In previous findings, regular nonwords not only produced a higher FA rate than did other items, but also a higher hit rate. In contrast, the present experiment showed that natural words and regular nonwords produced comparable hit rates which were both significantly higher than that yielded for irregular nonwords. Indeed, numerically, the hit rate was actually *higher* for natural words than regular nonwords. As will be shown in later experiments, this hit-rate advantage for natural words over regular nonwords appears to be a reliable one, and therefore contradicts previous findings on the hension effect. The significance of this hit rate pattern will be explicated further in Chapter 4 (see section 4.3)

The chief finding from Experiment 1, at least in relation to the discrepancy-attribution hypothesis, was that feedback exerted only limited influence on recognition judgments in the hension effect paradigm. Two conclusions could be made on this result. First, one could assume that the feedback manipulation was indeed *effective* in eliminating (and creating) perceptions of discrepancy, and yet the hension effect was nonetheless generated. In this case, the suitability of the discrepancy-attribution hypothesis in explaining the hension effect would be cast into doubt. Alternatively, one could conclude that feedback was *ineffective* in manipulating discrepancy. Because discrepancy was not eliminated for regular nonwords (and created for irregular nonwords), the hension effect was reliably found

in the false-feedback condition. Consistent with this view, it could be argued that the ineffectiveness of feedback might be due to the abstractness of its format and the subtlety of the false feedback manipulation. Because feedback was given in units of milliseconds, it might have been difficult for participants to determine what pronunciation duration would constitute fluent, as opposed to nonfluent, processing. Moreover, the size of adjustment made to the actual pronunciation duration in the false feedback condition was only 350 ms. In comparison to actual pronunciation duration data (see Table 1), it becomes clear that this 350 ms departure falls approximately within two standard deviations of the mean for regular nonwords, and well within one standard deviation of the mean for irregular nonwords. Thus, it would be reasonable to suggest that the degree of manipulation involved for the false feedback was too small for participants to notice that processing was nonfluent for regular nonwords and fluent for irregular nonwords. Given these reservations regarding the effectiveness of the feedback manipulation in Experiment 1, it might be that effects of feedback on perception of discrepancy (and hence recognition judgments) could still be demonstrated if a more concrete and meaningful form of feedback was implemented. Experiment 2 was carried out to test this hypothesis.

2.3 Experiment 2: Manipulating the Evaluation of Fluency (Descriptive Feedback on Speed)

Findings from Experiment 1 suggested that the pattern of FA rates in the hension effect might be impervious to feedback manipulations. However, it was also argued that the nature of the feedback, given in terms of pronunciation duration in milliseconds, might be too abstract as a fluency measure to be evaluated by participants in a meaningful way. Hence, a more concrete and meaningful indication of one's processing fluency is needed, and this could take form as labels (e.g., "fast", "average", "slow") which would describe the speed of one's processing for a stimulus. These descriptive labels were implemented in Experiment 2 in an attempt to modify the degree of discrepancy associated with items in the hension effect paradigm. In so doing, predictions made on the basis of the discrepancy-attribution hypothesis, in relation to FA rates, could be examined.

Unlike Experiment 1, participants in Experiment 2 were given feedback pertaining to the test item's processing fluency on every trial in the test phase.

Specifically, participants were (falsely) informed that their pronunciation duration for each item would be recorded, and promptly compared to a "group" average that was calculated from the data of others who have already taken part in the experiment. They were also told that if their pronunciation was slower or faster than this average, they would be informed accordingly with the feedback "slow" and "fast". If their pronunciation duration was approximately the same as the average, they would be told that their pronunciation duration was "average". In reality, these test instructions were misleading because no actual comparison was made between the participant's pronunciation duration and the group average. The type of feedback label presented on a given trial was randomly determined, and hence was not contingent on the participant's actual pronunciation duration for the item. For each item category (natural words, regular nonwords, irregular nonwords), one third of the items was assigned to one of the three ("fast", "average", "slow") feedback conditions.

Despite these procedural changes from Experiment 1, the objective of Experiment 2 remained the same. The feedback manipulation was employed for the purpose of altering the perception of discrepancy formed for items in the hension effect paradigm. If participants were able to apply feedback information strategically, the evaluation of an item's processing fluency would be modified, and under certain conditions, this would eradicate, or generate the perception of discrepancy. Specifically, discrepancy normally experienced with regular nonwords should be eliminated with the "slow" feedback label, because nonfluent processing would be consistent with the item's nonword status. Conversely, discrepancy would be created for irregular nonwords if these items were given "fast" feedback, because fluent processing would be discrepant to the subsequent failure in retrieving meaning for these items. Thus, it was predicted that the FA rate would be lower for regular nonwords in the "slow" feedback condition than for irregular nonwords in the "fast" feedback condition, which would essentially constitute a *reversal* of the hension effect. More generally, if externally-presented, fluency-related information was being consciously integrated by participants into their recognition process, feedback effects would be manifested in an increase in FA rates under the "fast" feedback, relative to the "average" and "slow" feedback conditions. The absence of a feedback effect, however, would suggest that recognition judgments are only dependent on subjective

assessments of fluency, and that this subjective evaluation is not susceptible to objective, external indicators.

2.3.1 Method

Participants. Eighteen University of Southampton undergraduates and postgraduates participated in this experiment either on a voluntary basis or in return for course credits. For all participants, English was the only language they could speak fluently, and none had participated in Experiment 1.

Materials and Design. Experiment 2 utilised the same materials as those in Experiment 1, but a different design was deployed. In each category (i.e., natural words, regular nonwords and irregular nonwords), the 60 items were divided into 2 sets of 30 items – one set of 30 items were "old" (i.e., presented at study and at test), the other 30 items were "new" (i.e., presented at test only). At test, each of these sets of old and new items was further divided into 3 subsets of 10 items each – with one subset assigned to each of the following feedback conditions: "fast", "average", and "slow". For the six practice test items, two items were also assigned to each of the three feedback conditions. Thus, excluding the practice items (three were presented at the beginning of study, and six at the beginning of test), the study phase consisted of 90 items (30 from each category) and the test phase consisted of 180 items (60 from each category). Of the 60 test items from each category, 20 items (10 old, 10 new) were given "fast" feedback, 20 items (10 old, 10 new) "average" feedback and the remaining 20 items (10 old, 10 new) "slow" feedback. Counterbalancing ensured that across participants, each item from each category was presented an equal number of times as old and new, and in the fast, average and slow feedback conditions.

Procedure. Participants were tested individually in a quiet cubicle. The procedure for the study phase was identical to that in Experiment 1 (see section 2.2.1). Following the study phase, participants were given their recognition test instructions. These were similar to those used in Experiment 1. That is, they were informed that their first task requirement for each test item was to read it aloud, and that the time taken for the pronunciation would be recorded. Unlike in Experiment 1, however, participants here were told that a number of people have already been tested and hence the "group average" for each test item's pronunciation duration was known.

Participants were further explained that their pronunciation durations for each item would be compared to this group average for the item, and their speed of pronunciation for the item would be labelled accordingly as "fast", "average", or "slow". That is, if the item was pronounced more quickly than the group average, the feedback label "fast" would ensue. If the pronunciation duration approximated the group average, the feedback would read "average". If the item was pronounced more slowly than the group average, the feedback would read "slow". To ensure that participants understood the meaning of the feedback, it was emphasised that because the "group average" of each test item was calculated on the basis of other participants' performance, there might be instances when the pronunciation duration might feel fast, but the feedback would still read "slow" if the duration was indeed longer than the group average. Likewise, there might be occasions when pronunciation might feel slow, but if it was indeed quicker than the group average, the "fast" feedback label would be given. In additional instructions, participants were told to read out the feedback (i.e., "fast", "average" or "slow") once it was presented, in order to demonstrate that they had registered the feedback. Following the feedback, they would then perform the old/new recognition judgment on the same test item.

Like in Experiment 1, each test item was presented first with an accompanying label "Read Word Aloud" located near the top of the screen, and the pronunciation duration for that item was recorded via a button press on the experimenter's response pad. However, unlike Experiment 1, after the pronunciation duration had been recorded, the test item was replaced by the feedback label ("fast", "average" or "slow") which was presented at the centre of the screen, in red and in uppercase. The duration of the feedback was 2 s, which was sufficiently long enough for participants to read this feedback aloud. Following the feedback, the test item was presented again, this time with the prompt "Old or New?" near the top of the screen. The participant's recognition judgment was given verbally to the experimenter, who then recorded this judgment using the response pad. The order of test item presentation was randomised with the constraint that a particular feedback ("fast", "average", or "slow") would be given for no more than four consecutive test trials. Apart from these changes, the procedure in Experiment 2 followed that in Experiment 1 (see section 2.2.1).

2.3.2 Results

Pronunciation Duration. Pronunciation duration data are presented in Table 4. As in Experiment 1, a 3 (item: natural word/ regular nonword/ irregular nonword) x 2 (status: old/ new) repeated-measures ANOVA was conducted on the pronunciation duration data to ascertain the effect of item type and previous study on the speed of pronunciation. This analysis showed a significant main effect of item, $F(2, 34) = 423.62, p < .001, MSE = 8639.97, \eta^2 = .961$. Post-hoc paired-samples t tests ($\alpha = .0167$) revealed that natural words ($M = 1041$ ms) were pronounced more quickly than regular nonwords ($M = 1092$), $t(17) = 6.92, p <. 001, SE = 7.41, \eta^2 = .738$. In turn, regular nonwords were pronounced more quickly than irregular nonwords ($M = 1617$), $t(17) = 21.06, p < .001, SE = 24.92, \eta^2 = .963$. Neither the main effect of status nor the item x status interaction was significant, $F < 1$, and $F(2, 52) = 1.21, p > .30$ respectively.

Table 4. Experiment 2: Means (and standard deviations) of pronunciation durations (in ms) for old and new items in each item type ($N = 18$).

	Old		New	
Natural	1043	(80)	1038	(76)
Regular	1089	(94)	1095	(87)
Irregular	1626	(189)	1607	(162)

Hit Rate. A 3 (item: natural word/ regular nonword/ irregular nonword) x 3 (feedback: fast/ average/ slow) repeated-measures ANOVA was performed on the hit rates obtained for the three item types under the three feedback conditions (means are presented in Table 5). This analysis produced a main effect of item, $F(2, 34) = 18.78, p < .001, MSE = .049, \eta^2 = .525$, reflecting the fact that hit rates were higher for both natural words ($M = .77$) and regular nonwords ($M = .73$) than for irregular nonwords ($M = .53$), $t(17) = 5.64, p < .001, SE = .043, \eta^2 = .651$ and $t(17) = 4.57, p < .001, SE = .044, \eta^2 = .551$ respectively for the two post-hoc comparisons ($\alpha = .0167$). There was no difference between the hit rates of natural words and regular nonwords, $t(17) =$

1.06, $p > .30$, $SE = .040$, $\eta^2 = .062$. Neither the feedback main effect nor the item x feedback interaction was significant, both Fs < 1.

Table 5. Experiment 2: Mean hit rates obtained for the three item types, in the three feedback conditions ($N = 18$). Standard deviations are in parentheses.

	Fast	Average	Slow
Natural	.76 (.14)	.78 (.14)	.77 (.18)
Regular	.75 (.16)	.72 (.15)	.71 (.22)
Irregular	.56 (.19)	.48 (.21)	.53 (.20)

FA Rate. FA rates produced for the three item types under the three feedback conditions (see Table 6) were also subjected to a 3 (item: natural word/ regular nonword/ irregular nonword) x 3 (feedback: fast/ average/ slow) repeated-measures ANOVA. This analysis revealed an item main effect, $F(2, 34) = 14.57$, $p < .001$, $MSE = .041$, $\eta^2 = .462$. This effect arose because more false alarms were produced for regular nonwords ($M = .46$) than for both natural words ($M = .26$) and irregular nonwords ($M = .31$), $t(17) = 5.81$, $p < .001$, $SE = .035$, $\eta^2 = .665$, and $t(17) = 3.26$, $p < .01$, $SE = .046$, $\eta^2 = .385$ for the respective comparisons. However, the difference in FA rates between natural words and irregular nonwords was not significant, $t(17) = 1.50$, $p > .15$, $SE = .035$, $\eta^2 = .116$ ($\alpha = .0167$ for these post-hoc t tests).

Although the overall effect of feedback on FA rate was nonsignificant, a second analysis was carried out to examine the effect of feedback more directly by comparing the "fast" and the "slow" feedback conditions only. This 3 (item: natural word/ regular nonword/ irregular nonword) x 2 (feedback: fast/ slow) repeated-measures ANOVA, as before, yielded a significant item main effect, $F(2, 34) = 12.05$, $p < .001$, $MSE = .032$, $\eta^2 = .415$. The central interest here, however, was that the main effect of feedback was significant, $F(1, 17) = 5.67$, $p < .05$, $MSE = .078$, $\eta^2 = .250$, indicating that the FA rate was higher in the fast ($M = .37$) than in the slow ($M = .32$) feedback condition.

Table 6. Experiment 2: Mean FA rates obtained for each item type, in each of the feedback conditions ($N = 18$). Standard deviations are shown in parentheses.

	Fast	Average	Slow
Natural	.27 (.12)	.28 (.19)	.22 (.15)
Regular	.48 (.17)	.48 (.21)	.42 (.18)
Irregular	.35 (.21)	.28 (.21)	.30 (.17)

As discussed earlier, a prediction based on the discrepancy-attribution hypothesis was that discrepancy would be eliminated if processing for regular nonwords is seen as nonfluent (i.e., if the feedback indicated slow pronunciation for these items). As a point of contrast, discrepancy would arise if the processing of irregular nonwords is seen as surprisingly fluent (i.e., if the feedback read "fast"). It followed then that a comparison between the FA rate for regular nonwords in the slow feedback condition and that for irregular nonwords in the fast feedback condition, might yield a result that departs from the usual hension effect pattern. This comparison was tested using a paired-samples t test, which showed that FA rates did not differ between regular nonwords in the slow feedback condition (.42) and irregular nonwords in the fast feedback condition (.35), $t(17) = 1.03$, $p > .30$, $SE = .070$, $\eta^2 = .058$. In contrast, the feedback would be consistent with general pronunciation speed, and with the assumptions underlying the discrepancy-attribution hypothesis, when regular nonwords were given "fast" feedback, and irregular nonwords "slow" feedback. The FA rates produced by these items, under these specific conditions, were also compared using a paired-samples t test. This test showed the typical hension effect pattern – significantly more false alarms were produced for regular nonwords in the fast feedback (.48), than for irregular nonwords in the slow feedback condition (.30), $t(17) = 4.06$, $p < .01$, $SE = .045$, $\eta^2 = .492$.

2.3.3 Discussion

Differences among the three item categories in processing fluency was again demonstrated in Experiment 2. Consistent with previous results, pronunciation

duration showed that natural words were processed more fluently than regular nonwords, which were in turn processed more fluently than irregular nonwords. However, no effect of item status was found. Contrary to previous findings from the repetition priming literature (e.g., Jacoby & Dallas, 1981), old items were not pronounced more quickly than new items. The reason for the absence of a repetition priming effect is unclear, although the failure to obtain the effect was not unprecedented in recognition memory research. Using pronounceable nonwords as their materials, Johnston et al. (1985, Experiment 2) also found that studied items and novel items did not differ in terms of speed of identification. In the current experiment, the participant's processing fluency was manually recorded by the experimenter. Hence, it could be argued that there existed a degree of measurement error which would obscure any subtle differences between the pronunciation durations of old and new items. Moreover, participants in the current study were not required to read the items aloud during study phase. Gains in fluency from prior experience with a stimulus might be more consistently achieved if the mode of processing was identical for study and test (cf. transfer-appropriate processing, e.g., Morris, Bransford, & Franks, 1977). Thus, if participants were also required to read the items aloud during study, a repetition priming effect, which shows faster pronunciation for old than for new items, might be more reliably obtained at test.

As in Experiment 1, the hension effect was successfully demonstrated in Experiment 2. Averaged across feedback conditions, substantially more false claims of recognition were made for regular nonwords than for natural words and irregular nonwords. Hit rate data, however, showed that natural words and regular nonwords produced similar hit rates, which were greater than that produced by irregular nonwords. This finding deviates somewhat from the typical hit rate pattern found by Whittlesea and Williams (1998, 2000), where regular nonwords generated higher hit rates than did natural words. The current data show this comparison to be in the opposite direction, with a slight advantage in hits for natural words than regular nonwords (see section 4.3 for further discussion on this finding).

The central interest of the present experiment, however, concerned the effect of feedback on FA rates. There was evidence that recognition performance on new items was influenced by the type of feedback given – across all items, more false

alarms were yielded in the "fast" than in the "slow" feedback condition. This finding suggests that participants were able to utilise fluency-related information in their decision making process during a recognition test. In relation to the discrepancy-attribution hypothesis, the purpose of the feedback manipulation was to alter the perception of discrepancy associated with the items. Specifically, it was predicted that discrepancy would be eliminated for regular nonwords in the "slow" feedback, and generated for irregular nonwords in the "fast" feedback condition. The corresponding comparison of the FA rates in these two conditions was nonsignificant, indicating that the FA rate difference between regular and irregular nonwords was diminished. However, it should be noted that this comparison was in the direction as stipulated by the hension effect – regular nonwords given "slow" feedback still produced more false alarms than did irregular nonwords given "fast" feedback (a 7% difference). This suggests that even though an objective, external measure of fluency exerted some impact on recognition judgments, the subjective assessment of processing ease remained the primary basis upon which fluency attributions were made.

The effect of feedback on the initial assessment of fluency, as found in Experiment 2, can be compared to the way that different sources of fluency were shown to be additive in their effects on the current perception of a stimulus (e.g., its pleasantness). For instance, previously studied words were judged to be more pleasant than novel words, but pleasantness ratings were further enhanced if the previously studied word was presented as the terminal word of a predictive sentence (Whittlesea, 1993, Experiment 5). In this example, both sources of fluency (from prior exposure to the item and from the context of a predictive sentence stem) were assessed through the subjective experience of item processing. In the current experiment, evaluation of fluency was principally based on the subjective perception of processing ease, but was shown to be enhanced by an externally presented indicator of processing fluency. In this, data from Experiment 2 demonstrated that participants were able to apply information relevant to the recognition judgment in a strategic and conscious manner.

2.4 Experiment 3: Manipulating Actual Processing Fluency

In relation to the discrepancy-attribution hypothesis, Experiment 2 provided some evidence that feedback, to an extent, could modify the level of discrepancy perceived for regular and irregular nonwords, and hence the FA rates produced for

these items. It could be argued, however, that this evidence was not overwhelmingly convincing. The specific comparison involving regular nonwords in the "slow" feedback and irregular nonwords in the "fast" feedback condition showed that the magnitude of the hension effect could be diminished to the point of statistical nonsignificance, but the effect was not reversed. The failure to achieve a reversal of the hension effect could be due to the way that the feedback manipulation was directed at the *evaluation* of processing fluency, rather than fluency per se. In this way, because processing fluency for an item had already been experienced, the effect of the feedback would be indirect and limited. Thus, the aim of Experiment 3 was to introduce a lexical decision task (LDT) requirement which would target processing fluency more directly, and in so doing, produce patterns of FA rates consistent with the predictions made from the discrepancy-attribution hypothesis. According to the hypothesis, it is the mismatch between processing fluency and subsequent performance (the retrieval of meaning) which creates a sense of discrepancy. Based on this reasoning, changes in processing fluency could determine the presence or absence of discrepancy, and hence the prevalence of false recognition.

In the original paradigm which generated the hension effect, participants were required to make three responses on each test item. In order, these responses were: pronunciation, lexical decision and recognition judgment (Whittlesea & Williams, 1998). In subsequent replications of the hension effect (e.g., Whittlesea & Williams, 1998, 2000), the LDT was omitted – participants only pronounced the test item and then made the recognition judgment. In this sense, pronunciation latency was held to be the indicator of fluency, and in turn the measure became the basis on which the discrepancy-attribution hypothesis was formulated. However, whereas the pronunciation latency data showed clear differences in fluency among all three item types, the response latency data from the LDT showed that response speed for regular and irregular nonwords did not differ from each other, but responding for both nonword groups were slower than for natural words (Whittlesea & Williams, 1998, Experiments 3 and 7). Whittlesea and Williams (1998) argued that because the LDT was the second task on the trial, response latencies obtained here could be influenced by the preceding pronunciation requirement. Hence, pronunciation latency, rather than lexical decision latency, would be a more appropriate index of fluency.

However, if lexical decision is the only response required from participants before the recognition judgment, LDT response latencies might indicate that regular nonwords are processed less fluently than natural words, and perhaps even irregular nonwords. Indeed, psycholinguistic research (e.g., Andrews, 1989, 1992; Coltheart, Davelaar, Jonasson, & Besner, 1977; see also Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) has shown that LDT response latencies were generally faster for real words than nonwords. Moreover, these studies have also found responding to be slower for nonwords with a high, rather than a low number of orthographic neighbours (i.e., real English words which differ from the nonword by only one letter; e.g., the word TENSION is a neighbour of HENSION). Recently, Cleary, Morris and Langley (2005) conducted a close inspection of the hension effect items provided by Whittlesea and Williams (2000), and found that regular nonwords, on average, have more neighbours than do irregular nonwords. On this basis, it was expected that regular nonwords would produce longer LDT response latencies than irregular nonwords.

Following this predicted pattern of LDT performance, discrepancy should not be experienced for regular nonwords because these items would not be fluently processed in the LDT. In contrast, discrepancy would arise for irregular nonwords if lexical decisions for these items are fluent. The elimination and generation of discrepancy, for regular and irregular nonwords respectively, would suggest that the hension effect would be reversed, such that the FA rate would be higher for irregular than regular nonwords. This prediction was tested in Experiment 3.

Although participants were not required to pronounce the test items in Experiment 3, the possibility remained that these items could nonetheless be pronounced (covertly) during the LDT, thus undermining the attempt here to manipulate discrepancy. Hence, to encourage that the LDT would be performed on the basis of orthography, a "250-ms" condition was devised such that the test items were only presented briefly, for a duration of 250 ms. The length of this duration was chosen because pronunciation latency data (e.g., Whittlesea & Williams, 1998) showed that participants required at least 800 – 900 ms before a pronunciation could be initiated. A presentation duration of 250 ms was therefore sufficiently short in preventing covert pronunciation, but sufficiently long for the item to be perceived

visually. A "clear" condition, where items remained on screen for both LDT and recognition responses was also included as a control condition in the experiment's design.

In both 250-ms and clear conditions, participants were instructed to give two responses in succession – first the lexical decision, and second the recognition judgment. If the opportunity for covert pronunciation was greater in the clear than in the 250-ms condition, the impact of the LDT on the perception of discrepancy might be diminished in the clear condition, relative to the 250-ms condition. Thus, the reversal of the hension effect (i.e., a higher FA rate for irregular than regular nonwords) might be more readily observed when items were presented for only 250 ms, rather than for an unfixed period in the clear condition.

Additionally, the LDT group as a whole would be compared to a no-task group whose participants were not asked to perform any task prior to recognition. In later replications of the hension effect, only the pronunciation requirement was preserved (e.g., Whittlesea & Williams, 1998, Experiments 5 and 6). Thus, another goal of Experiment 3 was to examine the necessity of a pre-recognition task in yielding the hension effect. If the existence of a separate pre-recognition task is critical in producing the hension effect, the no-task group would generate a pattern of recognition performance which differs from that observed by Whittlesea and Williams.

2.4.1 Method

Participants. Sixty-four psychology undergraduate students from the University of Southampton participated in return for course credits. All spoke English as their only fluent language and none had participated in previous experiments. Subsequent examination of the data revealed that one participant from the LDT group gave the same response (pressed the same button) for both the LDT and recognition judgment on every test trial. It was therefore reasonable to conclude that she had misunderstood the test instructions and therefore her data were discarded from the analyses. As a result, the LDT group contained 31 participants and the no-task group 32 participants.

Materials and Design. The essential difference between the LDT group and the no-task group was that at test, the LDT group made a decision on the test item's lexicality before giving the recognition judgment, whereas the no-task group were only required to make the recognition judgment. Thus, the materials and design of the experiment were identical for both participant groups. As in previous experiments, the materials consisted of three item categories (natural words, regular nonwords, and irregular nonwords), each containing 60 items. Each individual item was defined in terms of two factors – whether it would be old or new, and whether it would be presented in the 250 ms or clear condition during test. When these two factors were crossed, four counterbalancing conditions were created. Hence, in each category, the 60 items were further divided into 4 subcategories of 15 items each, with one subcategory assigned to each of the four following specifications: old-250ms, old-clear, new-250ms, new-clear. Counterbalancing ensured that across participants, each item appeared equally frequently as old and new, and in the 250-ms and clear condition. For each participant, there were in total 90 study items (i.e., 30 per category), and these 90 "old" study items were combined with the remaining 90 "new" items (i.e., 30 per category) to form the test phase, resulting in a total of 180 test items. At test, half (45) of the old (i.e., 3 x 15 per category) and half (45) of the new items (i.e., 3 x 15 per category) were presented in the 250-ms condition, and the remaining items (45 old, 45 new) in the clear condition. Because the presentation of the items in the 250 ms and in the clear condition was blocked, the order in which the blocks were presented was also counterbalanced across participants such that half of the participants received "250 ms" items first, and half received "clear" items first. Three of the six practice test items were presented at the beginning of the 250 ms block, and three at the beginning of the clear block. The presentation procedure of the practice items conformed with the block to which they were assigned.

Procedure. Participants were tested individually in a small room containing a PC. The procedure of the study phase was identical to that in Experiments 1 and 2. That is, after the standard study instructions, the practice items (in a fixed order), followed by the study items proper (in a uniquely randomised order) were presented to the participants, with each item presented for 1 s (see Experiment 1 in section 2.2.1 for more details). The test phase immediately followed the study phase, and here, separate pre-test instructions were given to the LDT and no-task groups. For the LDT

group, participants were told that the non-English words in both study and the following test phase were in fact "part of a new language that a psychologist is developing" and that these items were made of English letters but did not belong to the existing English vocabulary. The participants were then instructed, for each item of the recognition test, to respond "yes" if the item belonged to the "new language" (i.e., if it was a non-English word), and "no" if it was not (i.e., if it was an English word)[8]. This response was made by the participants by pressing the rightmost key ("yes") or the leftmost key ("no") on a response pad. Participants were told that following the lexical decision, they were required to make a recognition judgment on the same item. The no-task group received the same information concerning the new language words, but participants in this group were only told to make the recognition judgment for each item. Participants from both groups were instructed to give their responses by pressing the corresponding "old" (rightmost) and "new" (leftmost) button on the response pad.

For both LDT and no-task groups, the final part of the pre-test instructions detailed how items in the first half of the test phase would be presented, which, depending on the counterbalancing condition assigned to the participant, would either be in the 250-ms or clear condition. Specifically, participants who received the 250-ms block first were told that each test item would only be shown for a brief period of time, and that a row of asterisks (*****) would first appear at the centre of the screen to indicate where the item would be displayed shortly. They were warned to pay attention to this position cue lest they missed seeing the item. For participants receiving the clear block first, the instructions described that the item would be preceded by a row of asterisks to mark its position, and the item would remain on screen until a response was made.

---

[8] One might note that these instructions are the opposite of those usually found in LDT experiments, where participants are required to respond "yes" to real words and "no" to nonwords. The reason for instructing participants here to respond in a "reversed" manner was to further enhance the fluency experienced in responding to irregular nonwords, as "yes" responses have been found to be made more quickly than did "no" responses (e.g., Andrews, 1992). Due to their resemblance to real words, it was expected that lexical decisions for regular nonwords would be nonfluent, regardless of whether participants were instructed to respond "yes" (as in this case) or "no" to these items.

For the LDT group, each trial in the test phase began with a row of asterisks (*****) presented at the centre of the screen for 1 s. Following the asterisks, the test item was shown. At 250 ms after its onset, the test item was removed from the screen if it was in the 250-ms condition (but remained on the screen if it was in the clear condition). At the same instant, the labels "Right for YES" and "Left for NO" appeared on the right- and left-hand side of the screen respectively, and the label "New Language Word?" was placed at the top of the screen to remind participants of their LDT requirement. Once the LDT response was given, the LDT labels disappeared, and following a 500 ms blank, new labels – "Right for OLD", "Left for NEW", and "OLD or NEW?" were shown on the right-hand side, left-hand side, and top of the screen respectively. In the 250-ms condition, the test item did not reappear for the recognition judgment, whereas in the clear condition, the test item remained at the centre of the screen throughout both the LDT and the recognition components of the trial. The recognition judgment labels (and, in the case of the clear condition, the test item) were removed from screen once the recognition response was made.

For the no-task group, the procedure was identical, with the exception that the LDT component of the trial was omitted. Thus, each trial began with the presentation of a row of asterisks for 1 s, followed by the test item. In the 250-ms condition, this test item was visible on the screen for 250 ms only, whereas in the clear condition, the test item remained on screen until a response was made. At 250 ms after the onset of the test item, the recognition task labels were shown to prompt participants to make their recognition judgment. As in the LDT group, all labels (and, in the case of the clear condition, the test item) disappeared when the recognition response was made.

At the completion of the first test block of 90 test items, participants in both LDT and no-task groups were given the instructions for the second test block. They were told that for the second part of the test phase, their task remained the same (i.e., LDT group participants were to perform both LDT and recognition judgments, and the no-task group participants were to perform recognition judgments only). They were, however, informed that the presentation of the test items would be different, and depending on the counterbalancing condition, were given either the 250-ms or the clear presentation instructions. The second block was shown in the 250-ms condition for participants who received the clear condition first, and in the clear for those who

received the 250-ms condition first. The manner in which the test items were displayed (e.g., font, size, etc.) was the same as in previous experiments (see section 2.2.1 for details), and an ITI of 1 s was employed throughout the test phase.

2.4.2 Results

Two sets of data were produced in Experiment 3 – (a) LDT performance by the LDT group, and (b) recognition performance by both the LDT and no-task group.

LDT Data. The LDT performance by the LDT group was examined through two separate ANOVAs. In one, the dependent measure was response latency, that is, the elapsed time between the onset of the test item and the participant's response (means are shown in Table 7). In the other, the dependent variable was response accuracy, that is, the percentage of test items correctly judged as English or non-English (means are shown in Table 8). Both ANOVAs were repeated-measures, and had the following design: 3 (item: natural/ regular/ irregular) x 2 (presentation: 250 ms/ clear) x 2 (status: old/ new).

Table 7. Experiment 3: Means (and standard deviations) of LDT response latencies produced by the LDT group ($n = 31$), for the three item types in the two presentation conditions – 250 ms and clear. All data are in milliseconds.

	250 ms		Clear	
	Old	New	Old	New
Natural	1275 (463)	1281 (503)	1373 (453)	1381 (598)
Regular	1530 (639)	1522 (676)	1667 (630)	1760 (894)
Irregular	1281 (502)	1233 (557)	1516 (773)	1495 (739)

The analysis on response latency showed a significant main effect of item, $F(2, 60) = 15.30$, $p < .001$, $MSE = 196091.95$, $\eta^2 = .338$. Post-hoc comparisons with

Bonferroni adjustment to the alpha ($\alpha = .0167$) showed that averaged across presentation conditions and old/new status, response latency was significantly longer for regular nonwords ($M = 1620$ ms) than for both natural words ($M = 1328$), $t(30) = 4.29$, $p < .001$, $SE = 68.17$, $\eta^2 = .380$, and irregular nonwords ($M = 1381$), $t(30) = 5.05$, $p < .001$, $SE = 47.25$, $\eta^2 = .460$. Response latency was not significantly different between natural words and irregular nonwords, $t(30) = 1.05$, $p > .30$. There was also a significant main effect of presentation, $F(1, 30) = 5.30$, $p < .05$, $MSE = 559014.82$, $\eta^2 = .150$. Overall, responding was significantly faster for items in the 250-ms ($M = 1354$) than in the clear ($M = 1532$) condition. No other main effect or interaction was significant (highest $F = 2.59$, $p > .08$).

Table 8. Experiment 3: Means (and standard deviations) of LDT accuracy, in terms of percentage correct, as produced by the LDT group ($n = 31$) for the three item types in the two presentation conditions – 250 ms and clear.

	250 ms		Clear	
	Old	New	Old	New
Natural	94.62 (8.51)	94.62 (8.85)	97.20 (5.65)	97.20 (5.10)
Regular	83.44 (17.63)	86.02 (17.07)	83.44 (19.08)	84.95 (17.47)
Irregular	96.77 (5.13)	95.48 (5.81)	97.10 (5.15)	96.45 (4.63)

The second ANOVA on the response accuracy on the LDT showed a significant item main effect, $F(2, 60) = 19.33$, $p < .001$, $MSE = 294.20$, $\eta^2 = .392$. Post-hoc t tests ($\alpha = .0167$) revealed that averaged across presentation conditions and old/new status, lexical decisions were significantly more accurate for both natural words ($M = 95.91\%$ correct) and irregular nonwords ($M = 96.45\%$) than for regular nonwords ($M = 84.46\%$), $t(30) = 4.08$, $p < .001$, $SE = 2.81$, $\eta^2 = .357$, and $t(30) = 5.06$, $p < .001$, $SE = 2.37$, $\eta^2 = .460$ respectively. Accuracy did not differ significantly between natural words and irregular nonwords, $t(30) = .628$, $p > .50$. No other

significant main effect or interaction emerged from the analysis on the accuracy data (largest $F = 1.82, p > .15$).

Hit Rate. Recognition performance data (hit rates and FA rates) for both LDT and no-task groups are shown in Table 9 and 10 respectively. These two measures were analysed in separate 2 (group: LDT/ no task) x 3 (item: natural/ regular/ irregular) x 2 (presentation: 250 ms/ clear) mixed ANOVAs with group being the between-subjects factor and both item and presentation being the within-subjects factors. The analysis on hit rates revealed a significant main effect of item, $F(2, 122) = 19.14, p < .001$, $MSE = .044$, $\eta^2 = .239$. Post-hoc t tests ($\alpha = .0167$) indicated that this item effect arose because on average, natural words ($M = .68$) and regular nonwords ($M = .66$) produced a significantly higher hit rate than did irregular nonwords ($M = .53$), $t(62) = 4.81, p < .001$, $SE = .031$, $\eta^2 = .272$, and $t(62) = 6.28, p < .001$, $SE = .021$, $\eta^2 = .389$ for the respective comparisons. The hit rate for natural words did not differ from that for regular nonwords, $t(62) = .685, p > .45$. No other main effect or interaction was found to be significant (largest $F = 1.50, p > .20$). In particular, the group main effect, $F < 1$, as well as all interactions involving the group factor, were not significant (highest $F = 1.13, p > .25$), thus suggesting that the LDT and the no-task groups did not perform differently from each other.

Table 9. Experiment 3: Mean hit rates (and standard deviations) for LDT (n = 31) and no-task groups (n = 32), in the 250-ms and clear conditions.

	LDT				No-task			
	250 ms		Clear		250 ms		Clear	
Natural	.67	(.21)	.71	(.17)	.65	(.21)	.69	(.17)
Regular	.63	(.17)	.64	(.19)	.71	(.18)	.66	(.20)
Irregular	.49	(.17)	.52	(.22)	.56	(.22)	.54	(.16)

FA Rate. Similarly, the mixed ANOVA on FA rate data showed a significant item main effect only, $F(2, 122) = 25.10, p < .001$, $MSE = .029$, $\eta^2 = .292$. This effect

arose because averaged across groups and presentation conditions, regular nonwords (M = .35) produced significantly more false alarms than did irregular nonwords (M = .29), $t(62) = 3.30, p < .01$, SE = .019, η^2 = .150. In turn, the FA rate was significantly higher for irregular nonwords than natural words (M = .20), $t(62) = 3.63, p < .01$, SE = .025, η^2 = .175. As in the hit rate data, no other main effect or interaction was significant (largest $F = 2.72, p > .10$). There was no evidence that the LDT and the no-task groups produced different patterns of FA rates: $F < 1$ for the group main effect, and for an interaction involving the group factor, the highest $F = 2.13, p > .15$.

Table 10. Experiment 3: Means (standard deviations) for the FA rates in LDT (n = 31) and no-task groups (n = 32), in 250-ms and clear conditions.

	LDT				No-task			
	250 ms		Clear		250 ms		Clear	
Natural	.21	(.20)	.21	(.18)	.18	(.21)	.19	(.21)
Regular	.39	(.16)	.29	(.16)	.34	(.19)	.36	(.19)
Irregular	.30	(.17)	.26	(.13)	.31	(.16)	.28	(.15)

2.4.3 Discussion

As expected, the LDT response latency data indicated a different pattern of processing fluency from that seen in pronunciation latency (e.g., Whittlesea & Williams, 1998) or duration (see Experiments 1 and 2) data. In the LDT, regular nonwords were processed less fluently than were natural words and irregular nonwords. Accuracy data also conformed with this pattern – more errors were made for regular nonwords than for natural words and irregular nonwords. Conceivably, performance for regular nonwords was impaired by the way these items resembled real English words, thus making lexical decision difficult. The faster LDT response latency in the 250-ms condition, versus the clear condition, was an unanticipated finding. The exact reason for this main effect is unclear, although it might be the case that compared to the brief presentation in the 250-ms condition, participants took

advantage of the clear condition to examine the test item for longer, thus producing slower response latencies.

Turning to recognition performance, the hit rate data obtained here paralleled those observed in Experiments 1 and 2. Again, unlike Whittlesea and Williams (1998), there was a hit-rate advantage for both natural words and regular nonwords over irregular nonwords, and although not significant, the general trend was that natural words generated a higher hit rate than did regular nonwords. This finding will be addressed further in section 4.3 in Chapter 4.

Of greater relevance to the evaluation of the discrepancy-attribution hypothesis were the FA rate data. LDT response latencies clearly showed that processing was nonfluent for regular nonwords, and fluent for irregular nonwords. According to the discrepancy-attribution hypothesis, discrepancy should not be experienced for regular nonwords (since their nonfluency was consistent with their nonword status), but should be perceived for irregular nonwords (as their fluency was discrepant with their nonword status). It was therefore predicted that more false alarms would be produced for irregular than regular nonwords, thus reversing the hension effect. The replication of the hension effect in the LDT group therefore did not support this prediction.

However, an additional hypothesis made for Experiment 3 was that the presentation duration of the test item might influence the hension effect pattern produced by the LDT group. Specifically, it was argued that when participants have sufficient time to view the test item, a pronunciation could be covertly generated for the item, even if pronunciation was not a task requirement. The 250-ms condition was therefore included in order to prevent the covert pronunciation of the items, as this would reinstate the advantage in processing fluency for regular nonwords over irregular nonwords. Hence, it was predicted that the reversal of the hension effect might be more observable in the 250-ms condition, where pronunciation was prevented by the brief presentation (250 ms) of the test items, than in the clear-condition, where items were presented for an unfixed duration.

In view of this hypothesis, an additional 3 (item: natural/ regular/ irregular) x 2 (presentation: 250 ms/ clear) repeated-measures ANOVA was conducted to

specifically examine the FA rates produced by the LDT group. Along with an expected significant main effect of item (which indicated the presence of the hension effect), $F(2, 60) = 12.70$, $p < .001$, MSE = .021, $\eta^2 = .297$, there was also a significant main effect of presentation, $F(1, 30) = 4.30$, $p < .05$, MSE = .022, $\eta^2 = .125$[9]. The item x presentation interaction, however, was not significant, $F(2, 60) = 1.58$, $p > .20$. This indicated that the hension effect pattern was achieved regardless of the test item's presentation duration. This finding therefore contradicted the prediction that the hension effect might more likely be reversed in the 250-ms than in the clear condition. Indeed, numerically speaking, the hension effect (FA rate difference between regular and irregular nonwords) was actually larger in the 250-ms (9%) than in the clear condition (3%).

On the whole, the FA rate data from Experiment 3 indicated that the hension effect was generated regardless of the existence of a pre-recognition task, and that the effect was produced even when processing (as indexed by LDT response latencies) appeared to be more fluent for irregular than regular nonwords. A plausible explanation for the ubiquity of the hension effect across different experimental conditions is that even when pronunciation was not a task requirement, subvocalisation (or covert pronunciation) of the item was unavoidable. In support of this argument, recent work from Cleary et al. (2005) also showed that the hension effect was replicated when participants were not asked to pronounce the test item aloud, but only to perform a LDT on the item before the recognition judgment. Importantly, when subvocalisation was prevented through articulatory suppression (participants were instructed to verbally repeat "hi-ya-hi-ya" throughout the test phase), the FA rate difference between regular and irregular nonwords was eliminated. In view of Cleary et al.'s results, the presence of the hension effect in the 250-ms condition in Experiment 3 suggests that this method of subvocalisation

[9] This significant main effect of presentation was unanticipated, and it reflected the fact that in the LDT group, more false alarms were committed in the 250-ms (M = .30) than in the clear condition (M = .25). This effect might have arisen because of the unforeseen effect of presentation on LDT response latency – responding for the LDT was significantly faster (more fluent) in the 250-ms than in the clear condition (see Table 7). This enhanced fluency might in turn have increased the level of discrepancy perceived for regular and irregular nonwords. Consequently fluency misattribution was more likely to occur in the 250-ms condition, leading to the elevated levels of FA rates observed.

prevention (i.e., manipulation of presentation duration) might not have been effective. Together, these results also suggest that processing fluency might be primarily evaluated in terms of speed of pronunciation, which could have been performed either overtly or covertly.

2.5 Concluding Remarks for Chapter 2

Thus far, attempts to reverse the hension effect (specifically the FA rate difference between regular and irregular nonwords) have not been successful. This may suggest that the effect is unrelated to the way processing fluency is perceived to be discrepant, and that the discrepancy-attribution hypothesis is not applicable in explaining the elevated FA rate for regular nonwords. However, it could be argued that the manipulations so far targeted either the evaluation of fluency (Experiments 1 and 2), or actual fluency (Experiment 3), and hence, the failure in reversing the hension effect may simply be due to the way that the perception of fluency is not susceptible to external influences. An alternative argument might be that the manipulations employed so far have lacked the sufficient potency in modifying the level of fluency experienced. Either way, it might prove more fruitful to manipulate discrepancy through other means, for instance, by focussing on the items' meaningfulness. In the next chapter, a manipulation devised to modify items' meaningfulness will be described, and the findings of Experiment 4, which utilised this manipulation, will be reported.

Chapter 3

3.1 Experiment 4: Manipulating Items' Meaningfulness

As with previous experiments, Experiment 4 was founded on the premise that if the perception of discrepancy is associated with false recognition, the removal of discrepancy would therefore reduce false recognition. In the hension effect, the elevated FA rate produced for regular nonwords was argued to be a consequence of the discrepancy perceived during the processing of these items. Discrepancy was said to derive from the way that the initial fluency experienced for regular nonwords (as reflected by their relatively fast pronunciation latencies) was at odds with the subsequent failure to retrieve meanings for these items. According to this hypothesis then, discrepancy perceived for regular nonwords would be abolished if meanings could be retrieved for these items, and in turn, the hension effect would be eliminated.

In order for "meaning" to be successfully retrieved for nonwords, the manipulation imposed in Experiment 4 entailed the accompaniment of a "meaning label" to selected test items. In the case of natural words, the meanings would be veridical (e.g., "A material that hangs in a window" was the meaning for CURTAIN). However, in the case of regular and irregular nonwords, the meanings would be made-believe (e.g., "a style of Peruvian pottery", and "to sweeten a medicine with syrup" were the meanings for HENSION and STOFWUS respectively). To increase the level of credibility concerning the meaningfulness of these – in reality, nonsense – items, both pre-study and pre-test instructions would inform participants that some of the items are very rare English words, whose meanings are unlikely to be known to most people. Moreover, care was taken in the construction of these made-believe meanings, such that they did not correspond to meanings associated with existing English words.

It should be noted that the meaning labels were only applied at test. The meaning labels were not presented with items at study because this would allow deep, semantic encoding to occur (Craik & Lockhart, 1972), thus creating a depth of processing effect which would be irrelevant to the current investigation. More crucially, if meaning labels were provided at both study and test, participants could rely solely on the meaning labels in making their recognition decisions. That is, a

"meaningful" nonword test item could be rejected if the meaning label was judged to be novel. On these grounds, study items were presented individually, and for test items assigned to the "meaning" condition, the meaning label was given immediately prior to the test item's presentation. Recognition performance for these items would be compared to items in the "no-meaning" control condition whereby participants were not given the meaning of the to-be-presented test item, and were only informed of its word/nonword status.

As in previous experiments, predictions were formulated on the basis of the discrepancy-attribution hypothesis. For regular nonwords, discrepancy should be eliminated in the meaning condition, and thus, the FA rate for these items would be lower in the meaning than in the no-meaning condition. In contrast, it could be argued that for irregular nonwords, a variety of discrepancy, that of "surprising redintegration", might occur for these items (Whittlesea & Williams, 2001b). As mentioned earlier in the introductory chapter (see section 1.11), surprising redintegration was proposed to be the type of discrepancy perceived by participants in the "phraug" effect, where a high rate of false recognition was found for pseudohomophone nonwords like PHRAUG. Whittlesea and Williams (1998) argued that for these items, participants were first struck by the nonfluency in processing, but discrepancy was experienced because of the surprising success in retrieving a meaning which corresponded with the phonology of the pseudohomophone (namely, the meaning for "frog"). Similarly, in Experiment 4, irregular nonwords at test would be perceived as nonfluent initially, but under the meaning condition, the surprising success of acquiring a meaning for this nonfluent item would generate a sense of discrepancy (i.e., surprising redintegration). Consequently, it was expected that the FA rate for irregular nonwords would be higher in the meaning than in the no-meaning condition. Together, the predicted pattern of FA rates for regular and irregular nonwords would result in a diminishing, or even a reversal, of the hension effect in the meaning condition, in comparison to the no-meaning condition. That is, regular nonwords were expected to produce a similar, or even a lower FA rate than irregular nonwords when these two item categories were made meaningful at test.

3.1.1 Method

Participants. Twenty-four undergraduate and postgraduate students from the University of Southampton took part in this experiment in return for course credit or payment (£3). For all participants, English was the only language they could speak fluently. None had participated in previous experiments reported so far in this thesis.

Materials and Design. As in previous experiments, the full set of 60 items from each of the three categories (natural words, regular nonwords, irregular nonwords; see Appendix A) were used in Experiment 4. Additionally, short phrases pertaining to each item's meaning were constructed specially for the meaning condition. For natural words, these phrases were adapted from each item's real meaning as is defined in the Oxford English Dictionary (e.g., "A place where people wait for trains" – STATION). Sixty phrases were therefore constructed for the 60 natural words. On the other hand, meaning labels for nonwords were composed with the consideration that these descriptions did not correspond to meanings for existing English words. For example, the phrases "an ancient instrument shaped like a trombone" and "to whistle through one's teeth" were devised as meanings for the regular nonword BARDEN and for the irregular nonword LERTISP respectively. As only half of the nonword items (30 regular nonwords and 30 irregular nonwords) were presented in the meaning condition during test, only 60 of these "made-up" phrases were created to serve as meaning labels for the nonwords. It was ensured that for each participant, a particular meaning label corresponded only to one nonword. A full list of these meaning labels, along with their corresponding test items, can be found in Appendix C.

For the study phase, the 60 items from each category were divided into two sets of 30 – such that one set of 30 items from each category would be studied, thus making 90 study items in total. At test, all items from each category (i.e., 3 x 60 = 180 items in total) were presented. For the test phase, the 30 old and 30 new items from each category were further divided into 2 subsets of 15 items each, with one subset assigned to the meaning condition, and the other to the no-meaning condition. Thus, at test, each item category would contribute 15 old and 15 new items to the meaning and no-meaning condition respectively, giving a total of 90 test trials – 3 categories x (15 old + 15 new) – in each of the meaning conditions. Item assignment to old versus

new status, and to meaning versus no-meaning condition, was counterbalanced across participants such that each item was presented an equal number of times as old and new, and equally represented in the meaning and the no-meaning conditions. For the 6 practice trials presented at the beginning of the test phase, three were assigned to the meaning condition and three to the no-meaning condition.

Procedure. Participants were tested individually in a quiet cubicle. The standard study instructions, as used in previous experiments (see also Appendix B) were modified to inform participants that the study list would contain common English words, very rare English words, and non-English words. The reference "very rare English words" was intended for regular and irregular nonwords in the anticipation that some of these items would be presented in the meaning condition later at test. Because regular and irregular nonwords are in reality meaningless, participants were told that they might not know the meanings to these very rare English words. After these instructions, the study phase began with the three practice items, presented in a fixed order across participants, followed by the 90 study items, in a freshly randomised order for each participant. The presentation duration of each item was 1 s, and the ITI was also 1 s. In general, the procedure for the study phase was identical to that in previous experiments (see Experiment 1, section 2.2.1 for further details).

Immediately after the study phase, participants were given test phase instructions which contained the same references to common and very rare English words, and non-English words. These instructions also informed participants that on each test trial, a piece of information pertaining to the test item would first be presented, before the test item is given. On some trials, this information would indicate whether the following test item would be an English or a non-English word. On other trials, this information would consist of a description of the test item's meaning. Participants were instructed that for each trial, they were to first read aloud the given piece of information, and to press a designated "red" key on the response pad once they had finished reading. When the red key was pressed, the test item would then appear on the screen. Participants were told to read aloud this test item, and again press the red key after they had finished pronouncing it. This key press would produce instruction labels on the computer screen prompting participants to

make an old/new recognition judgment on the test item. Specifically, participants were to press the rightmost key on the response pad if they thought the test item was old, and the leftmost key if they thought the test item was new.

The six practice test trials were presented (in a fixed order across participants) at the start of the test phase, and these were followed immediately by the 180 test trials proper, which were presented in a freshly randomised order for each participants. On each trial in the test phase, the information pertaining to the test item was presented half way between the top and the centre of the screen. In the meaning condition, this information was a short phrase which described the "meaning" of the test item. For a natural word, this phrase was the real definition for the item; whereas for a regular or irregular nonword, this phrase was in fact a made-believe definition. In the no-meaning condition, the information preceding the test item would simply inform participants of its word/nonword status. This information was always veridical in that if the test item was a natural word, the label read "The following is an English word"; and for a regular and irregular nonword, the label read "The following is a non-English word".

After the information had been read by the participant, and the red key had been pressed, the test item was presented at the centre of the computer screen. When the participant finished pronouncing the test item, and had pressed the red key once more, the labels "Right if OLD" and "Left if NEW" would appear on the right- and left-hand side of the test item respectively. All labels and the test item were removed once the recognition response was made by participant via the response pad, and the next test trial commenced after an ITI of 1 s. As in previous experiments, participants were offered to take a short break halfway through the test phase (i.e., after 90 test trials). Other procedural details (e.g., the font, size, etc. of the presented items) were identical to those outlined in Experiment 1 (see section 2.2.1).

3.1.2 Results

Hit Rate. Mean hit and FA rates obtained for natural words, regular nonwords and irregular nonwords, in both meaning and no-meaning conditions, are shown in Table 11. A 3 (item: natural/ regular/ irregular) x 2 (meaning: meaning/no-meaning) repeated-measures ANOVA on hit rates showed a main effect of item, $F(2, 46) =$

11.89, $p < .001$, $MSE = .047$, $\eta^2 = .341$. This main item effect arose because the hit rate was significantly higher for natural words ($M = .69$) and regular nonwords ($M = .68$) than for irregular nonwords ($M = .49$), $t(23) = 4.07$, $p < .001$, $SE = .047$, $\eta^2 = .493$, and $t(23) = 4.74$, $SE = .038$, $\eta^2 = .569$ for the respective post-hoc paired-samples t test ($\alpha = .0167$). The hit rates for natural words and for regular nonwords did not differ significantly from each other, $t(23) = .21$, $p > .80$. Neither the meaning main effect, $F(1, 23) = 2.72$, $p > .10$, nor the item x meaning interaction, $F(2, 46) = 1.02$, $p > .35$ was significant.

Table 11. Experiment 4: Mean hit and FA rates for each item type in the meaning and no-meaning condition ($N = 24$). Standard deviations are in parentheses.

	Meaning		No Meaning	
	Hit	FA	Hit	FA
Natural	.72 (.17)	.31 (.23)	.65 (.19)	.25 (.19)
Regular	.70 (.20)	.45 (.20)	.65 (.19)	.36 (.14)
Irregular	.50 (.20)	.33 (.22)	.49 (.19)	.29 (.17)

FA Rate. As with the hit rate analysis, a 3 (item: natural/ regular/ irregular) x 2 (meaning: meaning/ no-meaning) repeated-measures ANOVA was performed on FA rate data. This analysis revealed a significant main effect of item, $F(2, 46) = 6.48$, $p < .005$, $MSE = .031$, $\eta^2 = .220$, reflecting the way that more false alarms were produced for regular nonwords ($M = .40$) than for natural words ($M = .28$) and irregular nonwords ($M = .31$), $t(23) = 3.20$, $p < .01$, $SE = .039$, $\eta^2 = .376$, and $t(23) = 3.05$, $p < .01$, $SE = .031$, $\eta^2 = .353$ for the respective post-hoc paired-samples t test ($\alpha = .0167$). There was no significant difference in the FA rate between natural words and irregular nonwords, $t(23) = .81$, $p > .40$. Apart from the item main effect, there was also a significant main effect of meaning, $F(1, 23) = 7.77$, $p < .05$, $MSE = .021$, $\eta^2 = .253$ which arose because averaged across item type, the FA rate was significantly higher in the meaning ($M = .37$) than in the no-meaning ($M = .30$) condition. The item x meaning interaction was not significant, $F < 1$.

3.1.3 Discussion

The significant finding from the present experiment was that the hension effect was replicated, regardless of whether regular nonwords were meaningful or meaningless. On the basis of the discrepancy-attribution hypothesis, it was predicted that discrepancy normally perceived with regular nonwords would be eliminated when the initial processing (pronunciation) fluency of these items was accompanied by the retrieval of meaning for these items. The provision of definitions in the meaning condition should therefore remove the sense of discrepancy for regular nonwords, and hence reduce the rate of false recognition. Contrary to these predictions, data from Experiment 4 showed that the FA rate for regular nonwords remained elevated in comparison to that of natural words and of irregular nonwords under the meaning condition – i.e., the hension effect was reliably observed. Indeed, if anything, the addition of meaning appeared to have boosted the effect, rather than reduced it. Although the item by meaning interaction was not significant, numerically speaking, the FA rate difference between regular nonwords and natural words was greater in the meaning (14%) than no-meaning (11%) condition. The same pattern was also evident in the FA rate difference between regular nonwords and irregular nonwords (12% and 7% in the meaning and no-meaning condition respectively).

For irregular nonwords, the provision of meaning was predicted to generate a sense of surprising redintegration because meaning was an unexpected outcome for these nonfluently processed items (Whittlesea & Williams, 2001b). As in the phraug effect (Whittlesea & Williams, 1998), which was also argued to be driven by surprising redintegration, it was hypothesised that the FA rate for irregular nonwords would increase under the meaning condition. Although the data seemingly supported this hypothesis, it was apparent, as evinced by the significant main effect of meaning, that FA rates were enhanced across all item types in the presence of meaning labels. Thus, there was no conclusive evidence that the increase in FA rates for irregular nonwords was necessarily a consequence of surprising redintegration. Rather, in some ways, the unanticipated finding that meaning increased FA rates resembled the "revelation effect" (e.g., Luo, 1993; Watkins & Peynircioğlu, 1990; Westerman & Greene, 1996, 1998). In the revelation effect, the probability of an "old" judgment for a recognition test stimulus is increased by requiring the participant to perform an

incidental task prior to the judgment. Typically, this incidental task may involve the stimulus itself (e.g., identifying the stimulus in its distorted form) or may be completely unrelated to the recognition item (e.g., solving an arithmetic problem prior to the recognition judgment on the word item). The data from Experiment 4 therefore suggest that relevant information, presented immediately before the test item was revealed, also produced an outcome akin to the revelation effect. In general, responses of "old" were more prevalent in the meaning than no-meaning condition, although this effect was only statistically significant in FA rates (but not in hit rates).

The effect of meaning in increasing FA rates may also be related to the concept of surprising coherence – a variety of discrepancy specified by Whittlesea and Williams (2001b). As delineated in the introductory chapter (see section 1.11), Whittlesea and Williams found an elevation of FA rates for words when they were presented as a terminal word for a predictive sentence stem, on the condition that the stem could be completed by a defined set of possible candidates (e.g., "She cleaned the kitchen with a ..." could only be sensibly completed with some kind of cleaning equipment like a MOP, BROOM, etc.), and that there was a pause before the terminal word was revealed. In the same way, the design of Experiment 4 might have created a sense of surprising coherence for natural words because the meaning label (e.g., "A place where people wait for trains") only corresponded to a set number of words (STATION, PLATFORM, etc.), and that there was a pause between the presentation of the meaning label and the test item (after reading the meaning, participants were to press a key to reveal the test item). However, although surprising coherence may explain how meaning inflated the FA rate for natural words, it is difficult to see how this concept could be used to account for the same effect observed for regular and irregular nonwords. For these items, the meaning labels were made-believe definitions with no-known English equivalents. Thus, it would be impossible for participants, before the nonword test item was presented, to generate a set of candidates which would correspond to the given definition. It was therefore unlikely that surprising coherence was experienced for regular and irregular nonwords, and was the underlying factor for the FA rate increase for these items in the meaning condition.

3.2 Comments and Recent Findings on the Discrepancy-Attribution Hypothesis

Collectively, the experiments reported so far in the current thesis have failed to provide evidence for the predictions made on the basis of the discrepancy-attribution hypothesis. Overall, the data from these four experiments pointed to two possible conclusions – either that the discrepancy-attribution hypothesis is not a viable account for the hension effect, or that the manipulations devised have not been effective in eliminating (or creating) perceptions of discrepancy. In regards to the latter case, there was an additional complication because no existing measure could conclusively verify that discrepancy had been successfully manipulated. In experiments carried out by Whittlesea and Williams (e.g., 1998; 2000), the presence or absence of discrepancy was inferred through fluctuations in FA rates alone. As argued earlier in the introductory chapter, the immeasurability of discrepancy poses a critical problem in the testing of the discrepancy-attribution hypothesis. It could be suggested that the experiments conducted so far were afflicted by the same problem – that the effectiveness of discrepancy manipulations could not be objectively substantiated.

Nonetheless, it remains difficult to see how discrepancy was not removed by the meaning manipulation imposed in Experiment 4. The manipulation did not target the evaluation of processing fluency (as in Experiments 1 and 2) or processing fluency itself (as in Experiment 3). Rather, the manipulation ensured that meaning was readily available when the regular nonword was being processed. Meaningfulness was therefore consistent with the item's fluent processing. Thus, there was no apparent reason as to why discrepancy was not successfully eliminated. In this way, the findings obtained from Experiment 4 may cause considerable problems for the discrepancy-attribution hypothesis as a tenable explanation for the hension effect.

As noted in Chapter 1, apart from the investigations carried out by Whittlesea and his colleagues, little research has emerged elsewhere which examined the discrepancy-attribution hypothesis or the hension effect. However, there have been two recent exceptions to this case. Earlier, the discussion to Experiment 3 (see section 2.4.3) had alluded to an experiment conducted by Cleary et al. (2005), which suggested that as long as subvocalisation is possible during the LDT, the hension effect would be produced. In another experiment reported in the same article, Cleary

et al. examined the generalisability of the hension effect to other materials. For example, these researchers constructed various objects (which were to be presented pictorially) on the basis of meaning and structural regularity. Three groups of objects were created – (a) meaningful and structurally regular (real-life objects such as a stool), (b) meaningless but structurally regular (structurally possible objects with no corresponding name in real-life), (c) meaningless and structurally irregular (structurally impossible and nameless objects). These three object types were analogous to natural words, regular nonwords, and irregular nonwords respectively. Cleary et al. found that the hension effect did not generalise to these materials. The two types of meaningless items, regardless of structural regularity, produced equally poor recognition performance (in terms of lower hit rates and higher FA rates), relative to the meaningful stimuli. Thus, unlike the letter-based materials in the original hension effect paradigm, meaningless but structurally regular objects (which were analogical to regular nonwords) did not produced an augmented FA rate in comparison to meaningless and structurally irregular objects (which were analogical to irregular nonwords).

Similarly, recent results from Reber, Zimmermann and Wurtz (2004) also questioned the applicability of the discrepancy-attribution hypothesis to hension-effect-based paradigms. Because their participants were Swiss-Germans, these researchers constructed their own set of natural words, regular nonwords and irregular nonwords in German. Moreover, instead of being presented in the context of a recognition memory task, these items were employed by Reber et al. in a duration judgment task. In many ways, this experiment resembled that reported in Whittlesea and Williams (1998, Experiment 7), where the hension effect item categories were presented for either 100 ms or 200 ms, and the participants were to judge whether the duration was "long" or "short". Whittlesea and Williams found that although processing fluency was greater for natural words than regular nonwords, and greater for regular nonwords than irregular nonwords (as indexed by pronunciation latencies), the proportion of items claimed to be "long", regardless of the actual presentation duration, was similar for both natural words and regular nonwords, and both were *higher* than that for irregular nonwords. Whittlesea and Williams (1998) again argued that the enhanced level of "long" judgments for regular nonwords was due to the

surprise fluency experienced for these items, and judgment of duration was not directly related to levels of processing fluency.

Reber et al. (2004), however, presented their three categories of German items at four different durations set between 32 ms and 80 ms, and participants were to judge the duration of the item on a 9-point (short to long) scale. They did not find evidence to support the discrepancy-attribution hypothesis, which predicted that regular nonwords, due to their surprising fluency, would be rated at least as long, if not longer, in duration than natural words. Rather, these researchers showed that participants appeared to base their duration judgments entirely on processing fluency. Regardless of actual duration, natural words were rated to be longer than regular nonwords, which were in turn rated to be longer than irregular nonwords. Reber et al. speculated that their failure to replicate Whittlesea and Williams's (1998) data might in part be due to the use of German materials, which might not be directly comparable to English-based stimuli.

Overall, Reber et al. (2004) also noted that the pattern of their duration judgments data did not follow that procured in recognition tests (e.g., Whittlesea & Williams, 1998). Based on the discrepancy-attribution hypothesis, duration would be rated as longest for regular nonwords, just as the response rate of old is highest for regular nonwords in recognition (Whittlesea & Williams, 1998). In view of the failure to obtain a similar hension effect in their duration judgment task, Reber et al. argued that the strategies involved in a duration judgment and a recognition judgment might be different. In the former, judgments could be made exclusively on the assessed fluency of the stimulus. On the other hand, recognition memory judgments may depend on recollection, with the effects of fluency being strategically discounted by participants (e.g,. Jacoby & Whitehouse, 1989). In short, processing fluency need not be considered by participants in making recognition judgments, whereas it is the critical factor which guides decisions regarding item duration.

Together, these above findings from Cleary et al. (2005) and Reber et al. (2004), as well as experiments described so far in this thesis, have failed to attain evidence which conformed to the predictions derived from the discrepancy-attribution hypothesis. In turn, it has become questionable as to whether the hypothesis is a suitable account for the hension effect. Reber et al.'s conclusion, in particular, hinted

at the possibility that recollection – one part of the dual-route perspective of recognition memory – may contribute to the explanation of the hension effect. Indeed, as will be outlined in the next chapter, recollection may be integral in explaining the hit rate pattern obtained in the hension effect paradigm. Furthermore, the following chapter will introduce the notion of memorability-based metacognitive strategies, and it will be argued that these strategies could be used to suppress items' FA rates, particularly in the case of natural words. In considering these factors, a more comprehensive account for the overall recognition performance observed in the hension effect may be formulated.

Chapter 4

In searching for a comprehensive account for both hit and FA data obtained in the hension effect paradigm, this chapter will systematically compare the recognition performance of regular nonwords with, first, irregular nonwords, and second, natural words. It will be argued that these two separate comparisons yield two distinct patterns of recognition performance. From these observations, it will be proposed that recognition judgments for nonwords are primarily driven by fluency-based familiarity, whereas for natural words, recollection plays a more important role in the recognition. Relevant to this latter proposition are the notions that fluency effects could be strategically discounted for natural words, and that perhaps the memorability of these items would encourage participants to use metacognitive strategies to suppress false alarm production. These two issues will be discussed in detail in this chapter. Additionally, Experiment 5, which concerns the memorability ratings of items in the hension effect paradigm, will be reported.

4.1 Recognition Performance Among Nonwords: A Concordant Pattern

In the previous two chapters, the reported experiments employed various experimental manipulations which were primarily directed at regular and irregular nonwords, rather than at natural words. In placing the focus on regular and irregular nonwords, an overall *concordant* pattern could be seen in the recognition performance of these two item groups. The term "concordant" is used here to describe the way that one item group is associated with both a higher hit rate and a higher FA rate than another item group in recognition (e.g., Maddox & Estes, 1997). In relation to the hension effect pattern, the results from Experiments 1 – 4 showed that among nonwords, both hit and FA rates increased as a function of orthographic regularity – the hit and FA rates were higher for regular nonwords than for irregular nonwords.

In other areas of research, concordant patterns in hit and FA rates are commonly found in investigations on the effects of word frequency on recognition memory. Experimentally, variations in linguistic frequency inherent in everyday words could be simulated in nonwords by controlling the amount of exposure allowed for these items in a pre-study "familiarisation phase". An example of this procedure is found in Maddox and Estes's (1997) investigation on the recognition performance of

nonwords (see also Chalmers & Humphreys, 1998; Dobbins, Kroll, Yonelinas, & Liu, 1998, Experiment 1; Greene, 1999). Prior to the study phase, participants in Maddox and Estes's experiment were presented with a list of nonwords in the familiarisation phase, where the presentation frequency of the nonwords varied between one and four. Some of these familiarised nonwords were presented again in a following study phase. In the final part of the experiment – a recognition test which required participants to answer "old" only to items presented in the study phase – Maddox and Estes found that hit and FA rates increased in accordance to the frequency of exposure in the familiarisation phase. Thus, greater exposure during familiarisation facilitated participants in recognising the item as old at test, if it had actually been presented during the study phase. However, greater familiarisation also misled participants into thinking that the item had been studied, when it had not actually been included in the study list.

The hension effect paradigm does not have a familiarisation phase during the experiment. However, it could be argued that differences in the nonwords' orthographic regularity constituted a manipulation equivalent to varying the amount of item exposure during the familiarisation phase. Although Whittlesea and Williams (1998) did not appear to have specific guidelines in constructing their regular and irregular nonwords, differences between these two groups, in terms of orthographic regularity, can be established more objectively using an index such as bigram frequency. Bigram frequency indicates how commonly a consecutive pair of letters within a given stimulus can be found in the English language, and this index has been widely implemented as a means of controlling orthographic regularity in psycholinguistic (e.g., Andrews & Scarratt, 1998; Peereman & Content, 1995) and verbal memory investigations (e.g., Farrell & Lewandowsky, 2003). For the hension effect materials, the mean frequency of all bigrams within each item was therefore calculated, and subsequently an analysis was carried out to determine any differences among item groups (see Appendix D for full details). It was found that indeed, regular nonwords are significantly higher in mean bigram frequency (M = 996.66) than are irregular nonwords (M = 531.72). This suggests that pre-experimental exposure, or familiarisation, is greater for the letter sequences in regular nonwords than in irregular nonwords. In this way, the concordant pattern observed for regular and irregular nonwords therefore aligned with previous findings of concordant recognition

performance as a function of items' familiarisation levels (e.g., Dobbins et al., 1998; Maddox & Estes, 1997).

In contrast, the recognition performance of natural words and regular nonwords do not form a concordant pattern. Experiments 1 – 4 demonstrated that while the hit rate of natural words was generally equal to, if not slightly higher than, that for regular nonwords, the FA rate was reliably lower for natural words than regular nonwords (this pattern constituted part of the hension effect). Defined in terms of item processing (e.g., pronunciation, or LDT response latency), natural words are more fluently processed than regular nonwords. Defined in terms of item characteristics (e.g., bigram frequency), the "familiarity" of natural words (mean of item-average bigram frequency = 924.85) and of regular nonwords (M = 996.66) are not significantly different from each other (see Appendix D for details). Hence, on the basis of these same measures, the FA rate for natural words should also be equal to, if not slightly higher than, regular nonwords.

4.2 Recognition Performance of Natural Words and Regular Nonwords: A Mirror Pattern?

In examining the non-concordant recognition performance between natural words and regular nonwords, it is notable that the hit rate pattern found for these two item types in Experiments 1 – 4 deviates somewhat from the typical outcome reported by Whittlesea and Williams (e.g., 1998, 2000). In a review of the experiments they had carried out on the hension effect, Whittlesea and Williams (2000) concluded that typically, more hits were produced for regular nonwords than for natural words and irregular nonwords. Thus, both the FA rate and the hit rate were enhanced for regular nonwords in the hension effect. Whittlesea and Williams (2000) further argued that this finding might reflect a more liberal response criterion being applied to the recognition judgments of regular nonwords. That is, participants were generally more inclined to respond "old" to regular nonwords than to natural words and irregular nonwords. The data from Experiments 1 – 4, in contrast, showed that hit rates obtained for natural words tended to be equivalent to, or in some cases exceed those for regular nonwords. Similarly, in the series of experiments conducted by Cleary et al. (2005) on the hension effect paradigm, the hit rate obtained for natural words was higher, rather than lower, than that produced for regular nonwords.

Thus, while Whittlesea and Williams's findings suggest that the recognition performance of natural words and regular nonwords follow a concordant pattern, the results from this current thesis, and from Cleary et al. (2005) suggest that the recognition data from these two item groups might form a pattern resembling more to the *mirror effect* (e.g., Glanzer & Adams, 1985, 1990; Glanzer, Adams, Iverson & Kim, 1993; Glanzer & Bowles, 1976). Simply speaking, the mirror effect describes the case where in a recognition test, one class of items produces a higher hit rate and a lower FA rate than another class of items. That is, performance is more accurate for one item class than the other (e.g., Glanzer et al., 1993). The most commonly-cited example of the mirror effect is that low frequency words produce a higher hit rate and a lower FA rate than do high frequency words – an effect also known as the *word frequency effect*, or WFE (e.g., Criss & Shiffrin, 2004; Hirshman & Palij, 1992; Malmberg & Murnane, 2002). In a similar way, because natural words in general produce a higher hit rate and a lower FA rate than regular nonwords, these two item categories can be said to produce a mirror pattern.

4.3 The Hit-Rate Difference Between Natural Words and Regular Nonwords: The Role of Recollection

To explain the mirror effect, an account based on the dual-route model of recognition memory (e.g., Jacoby & Dallas, 1981; Mandler, 1979, 1980) has been offered by Joordens and Hockley (2000). These authors argued that a mirror effect would arise when one class of items is "more familiar, but less recollectible" (p. 1550) than the other. In arriving at this conclusion, Joordens and Hockley imposed various manipulations (e.g., delay between study and test, speeded response during test) to reduce participants' reliance on recollection during a recognition memory task which involved low and high frequency words. They found that when the opportunity for recollection-based recognition was reduced, the hit rate portion of the WFE (that low frequency words produce a higher hit rate than high frequency words) was eliminated, or even reversed in some cases. Thus, it was argued that the typical hit rate pattern seen in the WFE was largely due to the way that low frequency words are more recollectible than high frequency words (see Guttentag & Carroll, 1997, for a similar view).

From the perspective of Joordens and Hockley's (2000) findings, it could be argued that in the hension effect paradigm, the suggested mirror pattern between natural words and regular nonwords reflects the way that natural words might be more recollectible than are regular nonwords. In particular, the trend for a hit-rate advantage of natural words over regular nonwords might be attributable to the greater involvement of recollection in the recognition of the former than the latter item group.

Perhaps somewhat surprisingly, Whittlesea and Williams (2000) concurred with the above proposition, and showed that natural words were more recallable than regular nonwords. In one of their experiments (Whittlesea & Williams, 2000, Experiment 1), these investigators presented participants with natural words and regular nonwords during study. In the recognition test, participants were required to rate each test item as one of the following: a) recall seeing the item, b) the item feels familiar, or c) the item is new. It was found that overall, the hit rate was higher for regular nonwords than natural words. However, there were more claims of recall for natural words than regular nonwords. Thus, the overall advantage in hits for regular nonwords over natural words was largely driven by the way that substantially more nonwords than natural words were claimed to be recognised on the basis of familiarity (for similar findings, see Greene, 2004). In contrast, the recognition of studied natural words was primarily driven by a recollection-based process.

Hence, despite evidence suggesting that natural words are more recallable, or recollectible than regular nonwords, Whittlesea and Williams's (2000) results still showed more hits were produced for regular nonwords than for natural words, a finding which was not observed in Experiments 1 – 4 here, or in Cleary et al.'s (2005) investigation. One plausible reason for this disparity in the findings could be due to the self-paced nature of the study phase in Whittlesea and Williams's experiments. Because the study procedure was self-paced, participants were able to control the presentation duration of each study item by allowing it to remain on screen until they struck a key to reveal the next item. This could be contrasted with the procedure used in Experiments 1 – 4 and by Cleary et al., where each item was studied for 1 s. Conceivably, the self-paced procedure could in some way enhance the recallability of regular nonwords if participants chose to spend more time to inspect these items. Thus, even though the overall recallability of regular nonwords was still less than that

for natural words when the study was self-paced, the difference in recallability between regular nonwords and natural words would probably be greater if both item types were exposed for the same durations. Hence, when the study duration was fixed, as in the current presents and Cleary et al.'s research, the contribution of recollection in producing hits would be substantially greater for natural words than for regular nonwords, leading to a, numerically-speaking, hit-rate advantage for the former over the latter item group.

4.4 Recollection-Based and Fluency-Based Recognition: Words versus Nonwords

To summarise, on the basis of the concordant pattern formed by regular and irregular nonwords, and the tentatively proposed mirror pattern formed by regular nonwords and natural words, it was postulated that recognition for nonwords is largely based on fluency-based familiarity, whereas recognition for words is primarily performed through a recollection-based process. Strong support for this conjecture can be found in Johnston et al. (1985), who demonstrated that the relationship between processing fluency and "old" responses in recognition was more direct for nonwords than for real words. For nonwords, perceptual fluency was found to be greater for false alarms than for misses, which suggests that participants were relying predominantly on perceptual fluency in forming their recognition judgments for these items. In contrast, as mentioned in Chapter 1 (cf. section 1.5), the relationship between perceptual fluency and the likelihood of "old" judgments for words was not as straightforward, with some old items being judged as new even when they had been fluently processed. It appears then that if test items, such as meaningful words, could be recognised on the basis of recollection or explicit memory, the contribution of fluency in recognition judgments would diminish. Conversely, in a follow-up investigation, Johnston, Hawley and Elliot (1991) showed that when the likelihood of recollection was decreased (e.g., when the encoding at study was shallow), a relationship between fluency and probability of "old" judgments was found, even for real word items. Overall, these findings from Johnston and his colleagues aligned with the suggestion that in the hension effect, recognition for words and nonwords is predominantly governed by recollection- and fluency-based processes respectively.

4.5 Recollection-Based Recognition: Strategic Discounting of Fluency

The above hypothesis, that recognition of natural words relies primarily on recollection, in turn implies that the high level of fluency experienced for natural words can somehow be ignored, or strategically discounted by participants. The findings from Jacoby and Whitehouse (1989), as described earlier in Chapter 1 (cf. sections 1.3, 1.12), may be relevant here. In the Jacoby-Whitehouse effect, judgments of old increased when the test item was preceded by a matching prime, in comparison to a nonmatching or control prime. Importantly, this effect was only found when the prime duration was short (16 ms) rather than long (600 ms), presumably because participants were aware of the prime's presence in the long, but not in the short prime duration condition. Jacoby and Whitehouse argued that aware participants were attributing the fluency experienced for the test item to the most plausible source, namely to the preceding prime, whereas unaware participants were attributing fluency to the past. A similar line of reasoning could be found in Higham and Vokey's (2000, 2004) account of the "identification heuristic", where participants infer their successful identification of a briefly presented test stimulus to its pastness. In Higham and Vokey's paradigm (see also Watkins & Gibson, 1988), each test item was first presented for a certain duration, participants would then attempt to identify it, and when the test item's true identity was revealed, participants were to make a recognition judgment on the item. It was found that when the initial presentation of the test item was short (50 ms), correct identification was associated with an increased probability of the item being judged as old. When the test item was initially presented for 250 ms, however, identification success did not translate to an increase of "old" judgments. Like Jacoby and Whitehouse, Higham and Vokey reasoned that identification success of the test item was attributed to its prior study when the presentation duration was short, but was attributed to the length of duration itself when the presentation was long.

The results from Jacoby and Whitehouse (1989) and Higham and Vokey (2000, 2004) are similar in the sense that the nature of current processing (be it fluency from a matching prime or accuracy in the test item's identification) was attributed to prior experience in one context but to item duration in another. Yet another interpretation of these findings is that in certain conditions, participants may

abandon their use of heuristics (based on either fluency or identification) in making their recognition judgments. In this view, in Jacoby and Whitehouse's paradigm, fluency experienced when the prime was long was not judged to be significant or *diagnostic* of an item's pastness. Consequently, participants might strategically discount the effects of fluency in deciding whether an item was old or new. Indeed, if anything, results from Jacoby and Whitehouse (1989) suggest that participants in such cases may make a misattribution of another kind. In the aware condition, where the prime duration was long (600 ms), the FA rate produced under the matching prime (.21) condition was significantly lower than that for both nonmatching (.36) and control prime (.33) conditions. It appears then when prime duration was made an obvious source of fluency, participants overcompensated for this processing ease in their discounting of its effects.

The notion that fluency can be strategically discounted in recognition decisions may be supported by research from Westerman and her colleagues (Westerman, Lloyd, & Miller, 2002; Westerman, Miller, & Lloyd, 2003; Lloyd, Westerman, & Miller, 2003), who found evidence that participants only used fluency as a basis of recognition if fluency was considered to be diagnostic of an item's pastness. For instance, the magnitude of the Jacoby-Whitehouse effect (i.e., higher levels of "old" responses for items following briefly-presented matching than nonmatching primes) was significantly greater when the modality of item presentation at study and test were the same (e.g., both visual) rather than different (e.g., visual at test, auditory at study; Westerman et al., 2002, Experiment 2). Moreover, when the modality of study items was manipulated on a within-subjects basis (i.e., half of the items were studied visually and half auditorily, and all items were tested visually), the magnitude of the Jacoby-Whitehouse effect was similar for visually-studied and auditorily-studied items. It appears then as long as a proportion of items were matched in modality between study and test, fluency-based recognition would take place on a global basis (i.e., for all test items). Together these findings suggest an appreciation on the participants' part that processing fluency would only be indicative of an item's pastness when the currently-encountered stimulus was presented in the same modality as when it was encountered earlier. However, it remained worthwhile to carry out fluency-based recognition judgements on a global basis even when fluency would be

diagnostic of an item's prior occurrence for only half of the time (Westerman et al., 2002).

The global manner in which the fluency heuristic was applied in recognition (Westerman et al., 2002) may appear to conflict with the postulation here that in the hension effect, recognition for nonwords, but not natural words, was based on fluency. In the same way that modality was a within-subjects manipulation in Westerman et al. (2002, Experiment 2), "item type" in the hension effect paradigm was also a within-subjects manipulation because each participant was presented with all three item classes. However, there are reasons to presume that participants in the hension effect paradigm would be able to selectively use fluency as a basis of judgment only for nonwords, but not natural words. In Westerman et al.'s (2002) within-subjects modality manipulation, half of the items were presented in the same modality (visual) in study and test, while the other half were presented in different modalities in study (auditory) and test (visual). In this design, in order for fluency to be used selectively (i.e., only for items presented in the same modality at study and test), participants would have to make an additional memorial judgment on whether the item was presented visually or auditorily at study. As this additional judgment would pose extra cognitive demand, it would therefore be likely that the participant bypassed this process, and consequently recognition judgments of all items would be based on fluency. In contrast, in the hension paradigm, the lexicality of an item (unlike the modality in which the item was studied) would be self-evident to participants at test, and thus, it is conceivable that participants would be able to utilise fluency only in the recognition of nonwords, but not natural words.

Elsewhere, evidence from Whittlesea and Price (2001) also suggests that the use of fluency in recognition judgments may be under participant's strategic control. These researchers showed that whether familiarity would be experienced when an old stimulus is encountered could depend on the approach being adopted in the processing of the study and test stimulus. In their experiment, processing fluency was found to translate into feelings of familiarity only when participants adopted a non-analytical approach to perceptual processing, where they were not specifically searching for a feature within a stimulus which would signify its "old" status. It was suggested that when a global form of perceptual processing was performed on a stimulus during

study, fluency would only arise when the same stimulus was also perceived in a global, non-analytical manner during test. If an analytical approach of perceptual processing was employed during test, where processing involved the search for a distinguishing feature within a stimulus, fluency, and subsequently feelings of familiarity would not be experienced.

From the perspective of Whittlesea and Price's (2001) findings, it could be proposed here that in the hension effect, an analytical approach involving the search of recollective details is adopted by participants in their recognition of natural words. The adoption of this analytical approach implies that the high level of fluency experienced for a natural word at test would be discounted when recollective details are being sought to indicate the item's prior occurrence. Consequently, recognition judgments for natural words would not be subjected to the effects of processing fluency. In contrast, a non-analytical approach would be associated in the recognition of nonwords. As a result, recognition for nonwords would primarily be based on fluency-induced feelings of familiarity, rather than on recollection.

4.6 Using Recollection to Suppress FA Rates: Recall-To-Reject Mechanisms

Although the studies detailed in the above section provide compelling evidence that fluency effects could be strategically discounted by participants, these studies did not specify the types of mechanism which could be applied to suppress FA rates. One of these mechanisms is the recall-to-reject process (e.g., Clark & Gronlund, 1996; Hintzman & Curran, 1994; Jones & Heit, 1993; Tulving, 1983), typically demonstrated in specialised recognition paradigms whereby the recall of a study item would allow the participant to deduce that the test item had not been studied previously. For example, in an associative recognition test (e.g., Odegard & Lampinen, 2005; Rotello, Macmillan, & Van Tassel, 2000), participants could reject an associative pair distractor (e.g., APPLE-CROWN) by recalling that the associative pair APPLE-DOG was presented in the study phase. (Participants were explicitly informed that a word item would not appear in more than one associative pair.) Similarly, in a changed-pluralisation task (e.g., Hintzman & Curran, 1994), items were studied either in their singular (e.g., TRUCK) or plural (e.g., TRUCKS) form, but never both. Thus, a distractor, which differed from the study item in terms of

plurality, could be correctly rejected if the participant was able to recollect its singular or plural form which had been presented during study.

Gallo (2004) used the term "disqualifying" to describe this type of recall-to-reject mechanism, because by recalling a particular item from the study phase, participants were able to use this evidence to disqualify the test item as a target. Further, Gallo demonstrated that participants could be encouraged to utilise this form of recall-to-reject process in the DRM paradigm (e.g., Roediger & McDermott, 1995). However, it was also shown that the strategy would be used only if the DRM lists were kept consistently short (i.e., each DRM list contained only three items). Presumably, short DRM lists increased the probability that all studied items could be recalled successfully, thus giving participants confidence to apply the recall-to-reject strategy. In contrast, partial recall of only some of the items in the DRM list did not lead to a reduction of falsely recognised critical lures (Gallo, 2004).

4.7 Using Memorability-Based Metacognitive-Strategies to Suppress FA Rates

Although this type of disqualifying recall-to-reject process is particularly effective in suppressing FA rates in certain experimental designs, it is unlikely that this mechanism would be responsible for the low FA rates achieved for natural words in the hension effect paradigm. Evidence gathered so far suggests that the use of the recall-to-reject strategy is restricted to specific recognition tests such as the associative recognition and changed-pluralisation tasks, where a distractor at test is in some way connected to a corresponding studied item. The recall-to-reject process could be adapted to the DRM paradigm (Gallo, 2004), but only if the study list was sufficiently short for all studied items to be recalled. Given these limitations, it is doubtful that recall-to-reject mechanisms could be applied for natural words in the hension effect, as these items are numerous in study and have little inter-item relatedness. Thus, some other mechanism must be responsible in aiding participants to correctly reject natural word lures, and in turn suppress the FA rate for this item group.

This other mechanism may entail the deployment of metacognitive strategies during recognition. In using metacognitive processes, participants could be described as actively searching for grounds to reject a test item as new, rather than accept it as old. An early example of how a metacognitive strategy could be used in suppressing

FA rates can be found in J. Brown, Lewis and Monk (1977). These researchers proposed the concept of "negative recognition", whereby items deemed to be highly memorable are associated with extremely low FA rates. *Item memorability* could be defined here as a subjective evaluation, made at the time of the recognition test, of how likely an event or stimulus would be associated with a clear memory had it occurred earlier (J. Brown, 1976). For highly memorable items, the absence of any recollective details could be inferred to be evidence of the item's novelty, and thereby a highly confident rejection could be made. In demonstrating memorability-based correct rejections, J. Brown et al. tailor-made each participant's recognition test according to the participant's personal details, which were assumed to be of high memorability. For example, in a recognition test consisting of first names, it was ensured that for each participant, his or her own name would serve as a lure. Across all participants tested, false alarms to the participants' own names never occurred. Further, these items were correctly rejected with extremely high confidence. J. Brown et al. argued that a highly memorable item, such as the participant's own name, could be confidently judged as new because memorial evidence was expected to be strong if the item had occurred earlier. Hence, the absence of any details retrieved for the item would strongly indicate that it had not been encountered before.

The impact of J. Brown et al.'s (1977) notion of negative recognition can be seen more recently in the research on memory for non-occurrences (e.g., Ghetti, 2003; Strack & Bless, 1994). The hypothesis underlying this area of research is that the memorability of a novel test item could influence the metacognitive strategy used by participants in making their recognition judgments. In turn, the type of strategy used would affect participants' tendency to endorse distractors (i.e., produce false alarms). The principles inherent in J. Brown et al's negative recognition follow most closely to the *"don't-recall-to-reject" strategy* – a strategy argued to be applicable when the test item is assessed to be highly memorable, but there is no clear recollection that the item had occurred earlier in study (Strack & Bless, 1994). In this scenario, the absence of recollection is taken to be evidence that the item had not been encountered in the past. Consequently, the participant would judge the item as new, because if such a memorable item had been presented earlier, it would be remembered (i.e., "If I saw it, I would remember it", Ghetti, 2003, p. 725). Additionally, Strack and Bless proposed the *"presupposition" strategy*, which was argued to operate when the test

item is evaluated to be unmemorable. In this case, although there is no clear recollection for the item, participants assume that they had forgotten its previous occurrence, and presuppose that the item had actually been presented before. As a result, the participant would judge the unmemorable item as old (i.e., "I must have forgotten it"). With the use of this strategy, the likelihood that new items in a recognition test are endorsed (i.e., the FA rate) would increase.

Evidence that these metacognitive strategies could be used by participants in recognition tests was reported by Ghetti (2003) and Strack and Bless (1994). In these experiments, participants were given pictorial stimuli of objects to study for a later recognition test. Most of these stimuli belonged to one semantic category (e.g., tools) and were therefore nonsalient because of their over-representation in the study phase. The remaining stimuli (i.e., the minority) were taken from another semantic category and therefore were of high saliency because they were less numerous in study. In the recognition test, it was found that highly salient distractors were less likely than nonsalient distractors to be misjudged as old. In addition, correct rejections were made with more confidence for highly salient distractors than nonsalient distractors (Ghetti, 2003). Moreover, when participants were led to believe that their recognition performance was impaired by the introduction of background noise, there was an increase in the FA rate for nonsalient distractors, whereas the FA rate for highly salient distractors remained low. These findings suggest that on the basis of the test item's assessed memorability, participants could accordingly adjust their tendency to endorse an item for which there was no clear recollection. If the item was assessed to be highly memorable, a don't-recall-to-reject strategy would be implemented and the FA rate would consequently be reduced. Conversely, the FA rate would be inflated if the presupposition strategy was used on the basis that the test item was considered to be unmemorable.

In view of Ghetti's (2003), and Strack and Bless's (1994) findings, it may therefore be worthwhile to speculate the involvement of metacognitive strategies in mediating the FA rate pattern of the hension effect. If natural words are considered to be highly memorable, then the don't-recall-to-reject strategy may be implemented in the recognition of these items. Consequently, a natural word distractor is likely to be rejected because in the absence of clear recollection, participants would assume that if they had seen it previously, they would have remembered it. As a result, the FA rate

for natural words would be suppressed. In contrast, nonwords are likely to be regarded as unmemorable and the presupposition strategy may be applicable for these items. In particular, because regular nonwords (unlike irregular nonwords) are reasonably high in processing fluency, participants might be especially inclined to presume the lack of recollective details for these items was a consequence of forgetting, and that these regular nonwords had actually been studied earlier. Consequently the use of this presupposition strategy would create an inflation in the FA rates for the fluent, but yet unmemorable regular nonwords.

4.8 Experiment 5: Measuring the Memorability of Items in the Hension Effect Paradigm

To begin the investigation of memorability-based correct rejections in the hension effect paradigm, Experiment 5 was conducted to ascertain the level of memorability associated with the three item types in the paradigm, as subjectively evaluated by participants. Previous attempts to measure memorability levels of recognition items have been carried out primarily in relation to the WFE (word frequency effect, e.g., Glanzer & Adams, 1985, 1990), as it had been hypothesised that memorability might underlie the FA rate difference between high and low frequency words (e.g., J. Brown et al., 1977). That is, the lower FA rate produced by low frequency words might be due to the way that these items are considered to be highly memorable, and thus compared to the less memorable high frequency words, stronger memorial evidence is demanded before low frequency words would be judged as old at test. Initial investigations on this hypothesis, however, showed the opposite outcome – participants generally rated high frequency words to be more memorable than low frequency words (e.g., Greene & Thapar, 1994; Wixted, 1992). It appeared then that memorability played no role in the production of the WFE.

Later investigations, on the other hand, suggested that memorability might, afterall, contribute to the generation of the WFE. Higher memorability ratings were obtained for low frequency than for high frequency words, when participants were asked to make these ratings in a "postdiction" context (Guttentag & Carroll, 1998). In the paradigm devised by Guttentag and Carroll, participants were given a standard recognition test consisting of high and low frequency words, and at test, memorability ratings were sought only for items that were judged as new. That is, for each item

receiving a "new" response, participants were asked to rate the likelihood of their recognising the item, *had* it been shown earlier during study. Guttentag and Carroll's results showed that memorability ratings gathered during an actual recognition test were vastly different from those gathered in other contexts, such as in a "mock" recognition test where none of the items were studied (e.g., Greene & Thapar, 1994; Wixted, 1992). Memorability ratings collected as postdictions conformed more closely to overall recognition performance – that is, low frequency words (associated with better performance) were rated to be more memorable than high frequency words (associated with poorer performance).

Guttentag and Carroll's findings were later replicated and extended by Benjamin (2003), who, within one experiment, demonstrated the shifting of memorability ratings for high and low frequency words. High frequency words were rated to be more memorable than low frequency words in the prediction context (i.e., when ratings were made prior to test during study), but the pattern was reversed – low frequency words were rated to be more memorable than high frequency words – when ratings were made for items judged as new during test. Further, Benjamin found evidence that through making postdiction ratings, participants appeared to have acquired the knowledge that low frequency words were more recognisable than high frequency words, and were able to utilise this knowledge in rating item memorability for a subsequent study list. In Benjamin's experiment, participants were first given a study list and a recognition test, and they were asked to make both predictions and postdictions of item memorability. Following this, they were given another study list for a second, hypothetical recognition test. Predictive memorability ratings for high and low frequency words in this second study list conformed with those gathered in the postdictive context. That is, low frequency words were judged to be more memorable than were high frequency words (Benjamin, 2003).

The above findings from Benjamin (2003), and Guttentag and Carroll (1998) therefore highlighted the importance of the context in which memorability ratings are collected. In view of this, this factor was considered and incorporated into the design of Experiment 5, where the aim was to obtain memorability ratings for items in the hension effect paradigm. Following Benjamin (2003), participants in Experiment 5 were required to rate the memorability for all items presented at study, and for test

items that were judged to be new. Additionally, after the test phase, participants were given an additional "post-test" list of study items (all previously unseen) for which they were to give memorability ratings. The purpose of obtaining post-test memorability ratings was to ascertain whether postdictive memorability ratings during test had an effect on participants' subsequent assessment of item memorability.

For the initial study phase (the "pre-test" phase), it was hypothesised that natural words would yield higher memorability ratings than would regular and irregular nonwords. According to research on the WFE, memorability ratings made prior to the test phase might be largely dependent on the amount of preexperimental experience participants have for the item. Because high frequency words are more commonly encountered in everyday life, they were rated (in the predictive context) to be more memorable than the less common low frequency words. In the same way, because natural words have been encountered preexperiemntally by participants, these items should be rated as more memorable than the previously unseen regular and irregular nonwords in the pre-test study phase.

However, if preexperimental exposure can be measured purely in terms of bigram frequencies, alternative hypotheses could be made for Experiment 5. Because natural words and regular nonwords do not differ significantly from each other in bigram frequency measures (see Appendix D), the pre-test memorability ratings for these two item groups should be equally high. Moreover, both natural words and regular nonwords would be rated as more memorable than irregular nonwords (whose bigram frequency measure is low) in the predictive context. Following Benjamin's (2003) findings, it was hypothesised that postdictive ratings collected during recognition test would differ from those gathered in the pre-test study phase. That is, at test, natural words would be rated as more memorable than regular and irregular nonwords, as recognition performance is more superior for real word than nonword items. This pattern of memorability ratings was also expected to be carried over to the following post-test study list, as participants have been shown to use their recognition performance in previous tests as a guide for their memorability ratings in subsequent study lists (Benjamin, 2003).

4.8.1 Method

Participants. The participants were 18 psychology undergraduate students from the University of Southampton. They took part in return for course credits. All spoke English as their only fluent language and none had participated in previous experiments described in earlier chapters.

Materials and Design. In order to create three separate phases: pre-test, test, and post-test, the 60 items from each category (i.e., natural words, regular nonwords, and irregular nonwords, see Appendix A), as used in previous experiments, were further divided into 3 subcategories of 20 items. From each item category, one subcategory of 20 items were presented in the first study phase (pre-test phase). These were mixed with another subcategory of 20 (designated as new items) and together, these 40 items were presented in the test phase. The remaining subcategory of 20 items were presented in the final part of the experiment – the second study phase (post-test phase). Three counterbalancing conditions were created such that each item served equally often as a pre-test study item, new test item, and post-test study item across participants. Thus, there were in total 60 items (20 from each of the three item categories) in the pre-test, and in the post-test study phase. The 60 items from the pre-test study phase were mixed with 60 new items (20 from each item category) to create a total of 120 test items. Finally, the six practice items (two from each category) used in previous experiments were again employed here as buffer items at the beginning of the study and test phases.

Procedure. Participants were tested individually in a quiet room. To ensure that they understood how to rate item memorability, participants were first introduced to the study-then-test structure of a typical recognition test. Specifically, a short example of a recognition test, which contained only three study items and six test items (all were words), was presented, and participants were told which items should be judged as "old" and which should be judged as "new" at test. Following this short demonstration, participants were given the instructions for the pre-test study phase. These instructions contained the standard set of details (see Appendix B), and informed participants that each study item would be presented for 1 s. Additionally, participants were told that after the presentation of each study item, they were required to rate the item on a 6-point scale, from "1 = I am sure I WILL NOT

recognise this item" to "6 = I am sure I WILL recognise this item". They were to give their ratings using the corresponding six buttons (labelled with the numbers from 1 to 6) on the response box.

After these instructions, the study phase commenced with the presentation of three practice study items (one from each category), followed by the 60 study items. These were presented in the manner as described in the method section in Experiment 1 (see section 2.2.1). Following the presentation of each study item (for 1 s), a rating scale, ranging from 1 to 6, appeared in black at the centre of the screen. The labels "I am sure I WILL NOT recognise this item" and "I am sure I WILL recognise this item" were placed directly under the numbers 1 and 6 respectively on this scale. An additional label "Please rate:" appeared above the rating scale to remind participants of their task. This scale and the accompanying labels remained on the screen until a response was given. An ITI of 1 s was deployed in between study trials.

At the completion of the pre-test study phase, participants were given instructions for the recognition test. They were asked to make recognition judgments for each test item by pressing the rightmost and the leftmost button on the response box for "old" and "new" judgments respectively. Unlike previous experiments, pronunciation was not required from participants. This modification in the procedure was imposed because results from Experiment 3 showed that the hension effect could be replicated despite the absence of the pronunciation requirement. The test instructions also informed participants that for every item they had judged to be new, they would be asked to rate the likelihood of recognising that item had it been studied earlier. Participants were told that this rating would be performed on the same 6-point scale employed in the pre-test study phase, but the labels were slightly modified to "1 = I am sure I WOULD NOT recognise this item" and "6 = I am sure I WOULD recognise this item". As a guide for this rating, it was suggested to participants that they could imagine seeing the item during study in another recognition test, and they could rate the likelihood of their recognising the item in that test.

After the instructions, the test phase commenced with the six practice items as used in previous experiments, followed by the 120 test items. Each test trial began of the presentation of the test item, along with the label "Old or New" at the top of the screen (acting as a reminder to participants of the first judgment required) and the

labels "Press RIGHT key if OLD" and "Press LEFT key if NEW" at the right of and the left of the test item respectively. The next test trial would follow (after a 1 s ITI) if the participant gave an "old" response. If the response was "new", however, the test item would be replaced by the 6-point rating scale, with the accompanying labels "I am sure I WOULD NOT recognise this item" and "I am sure I WOULD recognise this item" placed below the numbers 1 and 6 on the scale respectively. The label "Please rate:" was also presented above this scale to remind participants of their task requirement. Once a rating had been given, the next test trial was presented after a 1 s ITI.

At the end of the test phase, participants were informed of the requirement for the final part of the experiment (i.e., the post-test study phase), where they would be presented with another list of English and non-English words. It was specified that all of these items had not been presented in previous parts of the experiment. Participants were asked to perform the same task here as in the first study phase, that is, to rate each study item on its likelihood of being recognised in a later recognition test. They were told that this rating would be done on the same 6-point scale from the first study phase, with "1 = I am sure I WILL NOT recognise this item" and "6 = I am sure I WILL recognise this item". Following these instructions, the 60 post-test study items were presented. The format of item presentation and the rating procedure here was identical to that in the pre-test study phase.

4.8.2 Results

Two measures were of interest in Experiment 5 – recognition performance and memorability ratings. As in previous results sections, recognition performance was examined through separate analyses on hit and FA rates. For memorability ratings, two ANOVAs were conducted to ascertain any changes in the ratings obtained before and after test, and to analyse ratings collected during test.

Hit Rate. A one-way repeated-measures ANOVAs was conducted on the hit data (see Table 12), with item (natural/ regular/ irregular) as the within-subjects factor. The analysis on hit rates revealed a significant item main effect, $F(2, 34) = 25.64$, $p < .001$, $MSE = .015$, $\eta^2 = .601$. Post-hoc paired-samples t tests ($\alpha = .0167$) showed that the hit rate was significantly higher for natural words ($M = .92$) than

regular nonwords (M = .80), $t(17) = 4.55$, $p < .001$, SE = .027, η^2 = .549, and significantly higher for regular than irregular nonwords (M = .63), $t(17) = 3.80$, $p <$ .001, SE = .043, η^2 = .459.

Table 12. Experiment 5: Mean hit rates and FA rates for natural words, regular nonwords and irregular nonwords (N = 18). Standard deviations are in parentheses.

	Hit		FA	
Natural	.92	(.07)	.26	(.20)
Regular	.80	(.11)	.37	(.19)
Irregular	.63	(.20)	.24	(.16)

FA Rate. Similarly, the one-way repeated-measures ANOVA on FA rates (see Table 12), with item (natural/ regular/ irregular) as the within-subjects factor, produced a significant item main effect, $F(2, 34) = 5.06$, $p < .02$, MSE = .018, η^2 = .229. Post-hoc paired-samples t tests (α = .0167) indicated that the difference in FA rates between regular nonwords (M = .37) and natural words (M = .26) was marginally significant, $t(17) = 2.42$, $p < .03$, SE = .046, η^2 = .255, whereas regular nonwords produced a significantly higher FA rate than irregular nonwords (M = .24), $t(17) = 3.50$, $p < .01$, SE = .037, η^2 = .419. The FA rate difference between natural words and irregular nonwords was not significant, $t(17) = .40$, $p > .65$.

Memorability Ratings. Memorability ratings collected from the pre-test and post-test study phase, and from the test phase (for items which received a "new" response) are presented in Table 13. A 2 (context: pre-test/ post-test) x 3 (item: natural/ regular/ irregular) repeated-measures ANOVA was first conducted to elucidate potential differences in memorability ratings according to item type, and according to whether ratings were collected before or after test. This ANOVA revealed only a significant item main effect, $F(2, 34) = 126.14$, $p < .001$, MSE = .578, η^2 = .881. Neither the context main effect, nor the context x item interaction was significant, $F(1, 17) < 1$ and $F(2, 34) = 2.18$, $p > .10$, respectively. The significant item main effect was further scrutinised using post-hoc paired-samples t tests (α = .0167), which compared among the ratings obtained for the three item types, averaged across pre-test and post-test study phases. All comparisons were significant,

indicating that the average memorability rating was higher for natural words (M = 4.72) than for regular nonwords (M = 2.65), $t(17) = 12.26$, $p < .001$, $SE = .169$, $\eta^2 = .898$, and higher for regular than irregular nonwords (M = 1.99), $t(17) = 5.04$, $p < .001$, $SE = .131$, $\eta^2 = .599$.

Table 13. Experiment 5: Mean memorability ratings for natural words, regular nonwords and irregular nonwords obtained in the pre-test and post-test study phases, and during the test phase (N = 18). Standard deviations are in parentheses. Separate mean memorability ratings were calculated for "new" responses which constituted misses and correct rejections. Because the rate of misses and correct rejections differed across item types, the number of data points contributing to each cell average in the test phase is presented in square brackets.

	Study				Test			
	Pre-test		Post-test		Miss		Correct Rejection	
Natural	4.58	(.87)	4.86	(.75)	3.89	(.87) [28]	4.26	(.59) [207]
Regular	2.65	(.63)	2.65	(.58)	2.41	(.98) [56]	2.57	(.58) [173]
Irregular	2.03	(.58)	1.96	(.71)	2.32	(.74) [107]	2.14	(.62) [222]

Memorability ratings collected for items judged as new during test were analysed using a 2 (response: miss/ correct rejection) x 3 (item: natural/ regular/ irregular) repeated-measures ANOVA. The "response" variable refers to whether the ratings were produced following "new" responses to studied items (i.e., misses), or to non-studied items (i.e., correct rejections). The degrees of freedom in this analysis reflects that data from four participants were excluded because they produced perfect hit rates (1.00), and thereby did not produce any misses to be given memorability ratings. The overall ANOVA resulted in a marginally significant main effect of response, $F(1, 13) = 4.47$, $p < .06$, $MSE = .066$, $\eta^2 = .256$, because memorability ratings were in general higher for correct rejections (M = 2.99) than for misses (M = 2.87). The item main effect was significant, $F(2, 26) = 37.01$, $p < .001$, $MSE = .750$, $\eta^2 = .740$. Averaged across ratings for misses and correct rejections, natural words (M

= 4.07) generated significantly higher ratings than both regular nonwords (M = 2.49), $t(13) = 6.54$, $p < .001$, $SE = .241$, $\eta^2 = .767$, and irregular nonwords (M = 2.23), $t(13) = 6.65$, $p < .001$, $SE = .277$, $\eta^2 = .773$. The average ratings between regular and irregular nonwords did not differ significantly from each other, $t(13) = 1.65$, $p > .10$. Finally, the response x item interaction was not significant, $F(2, 26) = 2.13$, $p > .10$.

4.8.3 Discussion

Apart from the additional requirement to rate item memorability, the recognition test in Experiment 5 could be differentiated from those in previous experiments by having fewer number of items presented at study and test. Despite this procedural difference, however, the hension effect was again replicated. Although post-hoc comparison between regular nonwords and natural words was only marginally significant, the overall trend clearly showed that regular nonwords yielded a higher FA rate than did natural words and irregular nonwords.

When hit and FA rate patterns were jointly examined, a concordant pattern emerged from regular and irregular nonwords – regular nonwords produced both a higher hit rate and a higher FA rate than did irregular nonwords. As mentioned in previous sections, this concordant pattern is consistent with the proposal that recognition judgments for nonwords (regular or irregular) primarily rely on familiarity-based processes. Data on pronunciation latencies (e.g., Whittlesea & Williams, 1998) and durations (see Experiments 1 and 2, sections 2.2.2 and 2.3.2), as well as bigram frequency (see Appendix D), unequivocally suggest that greater fluency is experienced for regular than irregular nonwords. Thus, it appears that the fluency advantage of regular nonwords over irregular nonwords was directly translated to higher rates of "old" responses for regular nonwords at test, leading to a higher hit rate and FA rate observed for these items.

In contrast, a mirror pattern was formed by the hit and FA rates of regular nonwords and natural words. Previous experiments in this thesis have typically showed that the hit rate was higher for natural words than for regular nonwords. However, Experiment 5 was the first to demonstrate this hit-rate difference to be statistically significant. It is unclear why this comparison was not statistically significant in previous analyses. It might be argued that the significant outcome arose

in Experiment 5 because of the shorter study list and recognition test, compared to those in earlier experimental designs. However, a superior hit rate for natural words over regular nonwords was also obtained by Cleary et al. (2005), regardless of whether the study list was long (containing 90 items, Experiment 1c), or short (containing 60 items, Experiment 2a). Thus, the general trend of a hit-rate advantage for natural words over regular nonwords might be a reliable one.

For FA rates, the difference between natural words and regular nonwords is difficult to be reconciled by familiarity-based memory theories. Despite the fact that processing fluency of natural words is equal to, if not greater than, that of regular nonwords, the FA rate for natural words was substantially lower than that for regular nonwords. It was hypothesised that participants might regard natural words as more memorable than regular nonwords, and hence more compelling evidence would be required for an "old" response to be made for the former than the latter group of items. To garner support for this conjecture, the central purpose of Experiment 5 was to determine the participants' evaluation of the memorability of the hension effect materials. Given that the assessment of item memorability may vary across different stages of a recognition memory experiment (Benjamin, 2003; Guttentag & Carroll, 1998), memorability ratings were collected pre-test, during test, and post-test. Both pre-test and post-test ratings indicated that memorability ratings aligned with the wordlikeness or pre-experimental experience of the items. Natural words were rated to be most memorable, probably because they are meaningful real words and have been encountered pre-experimentally. Regular nonwords were judged to be less memorable than natural words, but more memorable than irregular nonwords, probably because while they are not real words, letter clusters found within these items are commonly encountered in everyday language (e.g., HENSION can be broken down into the elements HEN– and –SION). Irregular nonwords were seen as least memorable because they lack both meaning and are made up of uncommon sequences of letters.

Interestingly, the pattern of memorability ratings obtained during test differed slightly from those produced in pre-test and post-test study phases. For items judged as "new" at test, natural words were still regarded as the most memorable of the three item categories. However, there was no significant difference between the

memorability ratings of regular and irregular nonwords. Data from Guttentag and Carroll (1998), and Benjamin (2003) suggested that memorability ratings would be most predictable from actual recognition performance when ratings were made during recognition. Thus, it might be that memorability ratings did not differ between regular and irregular nonwords during test because of equivalent recognition performance for these two item groups. Unlike the mirror pattern, the concordant pattern does not indicate which of the two item groups forming the pattern is recognised more accurately than the other. In order to determine an index for recognition performance of each item class, a discrimination estimate, *d'* could be calculated based on hit and FA rates[10]. This *d'* estimate indicates the participants' ability to discriminate an old from a new stimulus at test. When estimates of *d'* were computed for each item group, and were analysed using a one-way repeated-measures ANOVA – with item (natural/ regular/ irregular) being the within-subjects factor – an item main effect emerged, $F(2, 34) = 20.82$, $p < .001$, $MSE = .299$, $\eta^2 = .551$. Post-hoc paired-samples t tests ($\alpha = .0167$) revealed that discrimination was significantly better for natural words ($M = 2.25$, standard deviation = .81) than both regular nonwords ($M = 1.30$, standard deviation = .66), $t(17) = 4.81$, $p < .001$, $SE = .197$, $\eta^2 = .577$, and irregular nonwords ($M = 1.17$, standard deviation = .54), $t(17) = 6.20$, $p < .001$, $SE = .174$, $\eta^2 = .693$. Importantly, regular nonwords and irregular nonwords did not differ significantly from each other in terms of *d'*, $t(17) = .75$, $p > .46$. This result is therefore consistent with the memorability ratings yielded during the test phase – regular and irregular nonwords were considered to be equal in memorability.

However, the analysis on pre-test and post-test memorability ratings yielded no evidence that there were any changes to the ratings post test. Thus, while the memorability ratings did not differ between regular and irregular nonwords at test (and thereby paralleling actual recognition performance), the ratings between these two item groups following test were, as in the pre-test phase, significantly different from each other. It therefore appears that unlike Benjamin's (2003) participants, memorability ratings established for regular and irregular nonwords during test were not transferred to the memorability ratings of a subsequent study phase. The reason

[10] In calculating *d'* for each participant, hit rates of 1 and FA rates of 0 were replaced by (1-1/2N) and (1/2N) respectively, where N was the maximum possible number of hits and false alarms which could be made in the condition (Macmillan & Creelman, 2005).

for this failure to replicate Benjamin's results is unclear, although it could be argued that the non-significant difference in recognisability between regular and irregular nonwords at test might be difficult for participants to detect from test. Put another way, numerically, the mean *d'* estimate (and mean memorability ratings) for regular nonwords was still higher than that for irregular nonwords. This trend was therefore consistent with pre-test ratings of item memorability. In comparison, in the WFE paradigm (Benjamin, 2003), the change in participants' preconceptions regarding item memorability was far more striking – participants originally misinterpreted that high frequency words would be more memorable than low frequency words, but the subsequent recognition test demonstrated the exact opposite trend, thereby prompting participants to reverse the initial, misconstrued pattern of memorability ratings for high and low frequency words.

At test, the response main effect approached significance ($p < .06$), suggesting that overall, correctly-rejected distractors were deemed to be more memorable than missed targets. A cursory examination of the data indicates that this effect might be primarily driven by natural words. A paired-samples t test, comparing the memorability ratings of misses and of correct rejections for natural words was marginally significant, $t(13) = 1.89$, $p < .09$, $SE = .196$. In view of the fact that four participants were eliminated from the analysis for natural words, the failure to obtain an outright significant t statistic may be due to the low N – that is, insufficient power. Hence, there was some evidence that correctly-rejected natural words were, on average, regarded as more memorable than missed natural words. This result suggests that a number of targets might have been rejected (mistakenly) because they were considered to be low in memorability. On the other hand, the majority (88%) of "new" responses were correctly given to distractors that were considered to be highly memorable. In contrast, for regular nonwords, the comparison between memorability ratings of missed and of correctly-rejected items was far from statistical significance, $t(17) = .78$, $p > .45$, whereas the same comparison was in the opposite direction for irregular nonwords (correctly-rejected items were rated as more memorable than missed items). Together, these findings may lend credence to the proposal that item memorability plays a particularly critical role in the correct rejection of natural word lures, but appears to have little influence on the correct rejection of regular and irregular nonword lures.

4.9 Concluding Remarks for Chapter 4

In this chapter, it was postulated that in the hension effect paradigm, recognition for regular and irregular nonwords is primarily driven by fluency- or familiarity-based processes, whereas recollection-based processes play an integral role in the recognition for natural words. For nonwords, a reliance on fluency in recognition judgments was argued to give rise to a concordant pattern of recognition performance, where regular nonwords, due to their greater fluency, produced higher hit and FA rates than did irregular nonwords. The pattern of recognition performance between regular nonwords and natural words, in contrast, was similar to that seen in the mirror effect – the hit rate was higher, but the FA rate was lower, for natural words than for regular nonwords. The role of recollection was implicated in the hit-rate difference between these two item groups. In relation to the FA-rate difference, it was suggested that the lower FA rate found for natural words was achieved through the application of a metacognitive strategy. By this, it was argued that natural words were evaluated to be highly memorable items, and consequently, when no memorial information could be retrieved for them, these items could be rejected on the basis that if they had been studied, recollective evidence would be readily available.

To begin the investigation on how memorability-based strategies could be used to reduce FA rates, Experiment 5 was carried out to provide data regarding the memorability, as assessed by participants, of the three item types in the hension effect paradigm. As expected, natural words were consistently rated to be the most memorable item group. Furthermore, memorability ratings at test paralleled closely to actual recognition performance, as indexed by the discrimination index *d'*. Here, memorability ratings for regular and irregular nonwords did not differ from each other, but both of these item groups were evaluated to be significantly less memorable than natural words.

The upcoming chapters will report several experiments which were conducted to explore the role of metacognitive strategies in the recognition of items from the hension effect paradigm. The question central to these experiments was whether item memorability, when modified through experimental manipulations, could in turn affect recognition performance, particularly in terms of FA rates. An additional aim of these following chapters will be to outline a signal-detection model which could

potentially account for the data yielded in the hension effect. In positing such a model, particular focus will be placed on the way a memorability-based rejection mechanism can be expressed in signal-detection terms.

Chapter 5

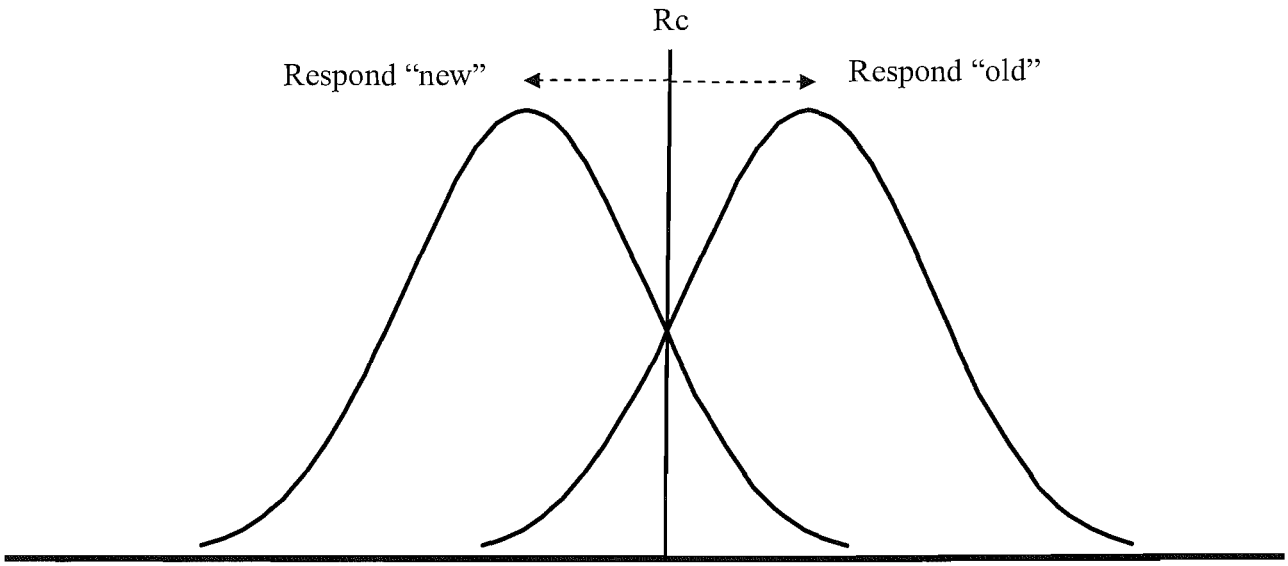
In the previous chapter, it was proposed that the high levels of memorability in natural words allowed participants to correctly reject a large number of lures from this item category, thus resulting in the low FA rate observed for natural words in the hension effect paradigm. In recognition memory research, the use of memorability in making correct rejections has been modelled by a oft-used framework based on *the theory of signal detection*. More generally, signal-detection (SD) models are widely regarded as particularly suitable in encapsulating recognition memory performance, because recognition memory tests are essentially tests of choice and decision making – on a recognition test, the participant's basic task is to judge whether an item is "old" or "new" (McNicol, 1972). The first aim of this chapter, therefore, is to introduce the basic principles of SD theory. Specifically, because it was argued here that recognition performance of natural words and regular nonwords form a mirror pattern, the application of SD theory in modelling the mirror effect will be described in detail. Furthermore, particular attention will be placed on how metacognitive processes, such as memorability-based rejection mechanisms can be represented in the model. In relation to this issue, experiments reported in the coming chapters will investigate the possibility of reducing the FA rates of the three item groups in the hension effect paradigm, through the experimental enhancement of item memorability. Findings from these experiments will in turn be interpreted from the perspective of SD models.

5.1 Signal-Detection Theory and Recognition Memory

The basic structure of the SD framework is depicted in Figure 1. In this framework, "old" items and "new" items in a recognition test are represented by two Gaussian distributions placed on an underlying, unidimensional continuum. In adopting SD theory to account for recognition data, different memory theorists have assigned their own labels for this continuum. For example, from the viewpoint of the dual-process model of recognition memory (e.g., Jacoby & Dallas, 1981; Mandler, 1980), this continuum is labelled as "familiarity", whereas other theorists have associated this continuum with decisional probability (e.g., Glanzer & Adams, 1985, 1990). In Figure 1, a less theory-laden term, "memory strength" has been adopted to label this underlying dimension. In a recognition test, because studied, old items (targets) are stronger in memory strength than are non-studied, new items (lures),

targets form a separate distribution which is located further to the right on the continuum than do lures. As mentioned in the previous chapter (see section 4.8.3), the discrimination estimate, *d'*, is an index of how well participants differentiate targets from lures. In the SD model, *d'* is represented by the distance (in standard deviation units) between the means of the target and lure distributions. Greater distance (and less area of overlap) between the distributions equates to better discrimination (Green & Swets, 1966; Macmillian & Creelman, 2005; McNicol, 1972).

Figure 1. SD model for recognition memory with distributions for targets and lures. The underlying dimension represents memory strength. Response criterion (Rc) determines whether response would be "old" or "new".



It is only with the addition of the response criterion, however, that a full picture of the participant's recognition performance emerges. An "old" response would ensue when an item's memory strength exceeds the point marked by the response criterion (Rc in Figure 1). To the right of the response criterion, the area bounded by the old distribution, and that by the new distribution, corresponds to the participant's hit rate and FA rate respectively. The placement of the response criterion, relative to the intersection of the target and lure distributions, can be expressed in terms of the statistic *C*, an estimate of bias (e.g., Macmillan & Creelman, 2005; Snodgrass & Corwin, 1988). Positive values of *C* indicates a conservative bias, such that the criterion is placed further to the right of the distributions' intersection,

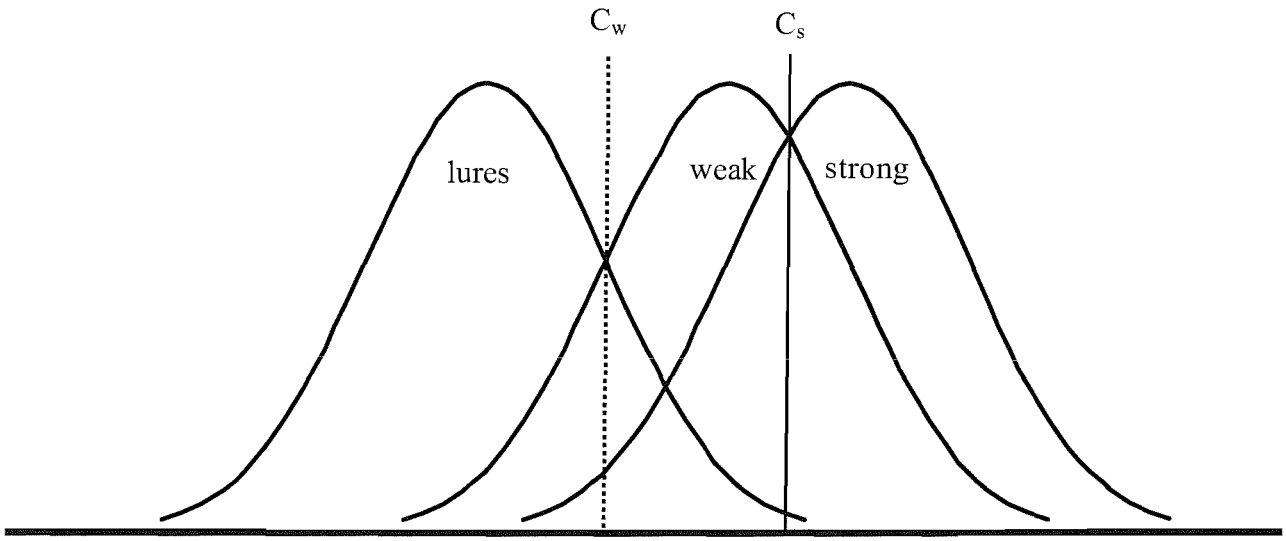
resulting in a lower FA rate, but at the same time, a lower hit rate. In contrast, negative values of C reflects liberal responding, such that the criterion is placed further to the left of the distributions' intersection, leading to an increase in the hit rate but also an increase in the FA rate. Optimally, the response criterion should be placed directly where the old and new distributions intersect (as in Figure 1, such that $C = 0$), as accuracy is maximised at this point.

5.2 A Criterion-Shift SD Account for the Mirror Effect

As introduced in the previous chapter, the mirror effect, such as that observed between natural words and regular nonwords in the hension effect paradigm, is produced when one class of items has a higher hit rate and a lower FA rate than the other. Again, as mentioned earlier (see section 4.2), a well-known example of the mirror pattern can be found in the word frequency effect (or WFE, e.g., Criss & Shiffrin, 2004; Hirshman & Palij, 1992; Malmberg & Murnane, 2002) – that is, recognition performance has been widely shown to be more accurate for low frequency than high frequency words.

One type of SD-based model which has been proposed to account for the WFE, or more generally, the mirror effect, is the *criterion-shift model* (see Figure 2). In this account, differences in hit and FA rates between different classes of items could be explained by assuming a unique response criterion for each item class. That is, for each individual test item encountered during the recognition test, the participant would adopt a response criterion that is specific for the item class to which the item belongs. In this model, the underlying continuum is typically given a label pertaining to memory strength such as "familiarity", rather than decisional probability. As in the labelling of the continuum in Figure 1, the generic label "memory strength" will be used hereafter to refer to the underlying dimension in criterion-shift models, such that ambiguous references to extant recognition memory theories can be avoided.

Figure 2. Criterion-shift account for the mirror effect in recognition memory. There are separate distributions for strengthened (strong) and non-strengthened (weak) targets, but only one distribution for the lures. The criterion used to respond to strong items (C_s) is more conservative than that for weak items (C_w).



An early version of the criterion-shift account can be found in J. Brown et al. (1977). As mentioned in Chapter 4, these researchers demonstrated the notion of "negative recognition", whereby correct rejection in recognition tests could be made on the basis of how memorable an item was assessed to be. For example, highly memorable items (such as the participant's own name) could be confidently judged as new if there was no evidence that it had occurred earlier. From the perspective of the criterion-shift SD model, J. Brown et al.'s (1977) theory of memorability-based rejections assumes that new items of high and low memorability amalgamate to form one distribution, but the distributions for strong targets (those high in memory strength and memorability) and weak targets (those low in memory strength and

memorability) are distinctly separate[11]. As Figure 2 shows, on the underlying continuum (representing memory strength), the distribution for strong targets is also placed further to the right of the distribution for weak targets. Moreover, in terms of distance from the new distribution, the strong target distribution is further apart than is the weak target distribution. The mirror pattern would arise when participants adopt a more conservative response criterion for memorable items than for unmemorable items. In other words, criterion placement is affected by item's memorability (J. Brown, 1976).

Likewise, in relation to the WFE, a criterion shift model was endorsed by Gillund and Shiffrin (1984), who argued that different response criteria are adopted by participants in creating the mirror effect with high and low frequency words. In these criterion-shift models, it is assumed that participants are able to distinguish items in terms of their memory strength or memorability, and set the response criterion accordingly. On this assumption, participants are expected to assess low frequency words to be more memorable than high frequency words. Along this line of reasoning, it could be suggested that in the hension effect paradigm, natural words are deemed to be more memorable than regular nonwords, and thus the response criterion is set more conservatively for natural words than for regular nonwords, thus resulting in the FA-rate component of the mirror pattern.

5.3 Mirror Effects Arising from Experimental Manipulation of Strength

A commonality between the WFE, and the mirror pattern seen in the hension effect, was that both phenomena were generated solely through differences in item characteristics that are *preexperimentally* determined (i.e., word frequency,

---

[11] It should be noted that the focus of the research by J. Brown and his colleagues (J. Brown, 1976; J. Brown et al., 1977) was on memorability as a mechanism for making correct rejections in a recognition test, and hence was relevant to lures. Following this framework, memorability is viewed here as a feature that is subjectively evaluated by participants when new items are assessed, and is subsequently used as a basis for determining the placement of the response criterion. In contrast, the term "memory strength" is used in relation to targets – strong targets (which are high in strength) can be differentiated from weak targets (which are low in strength). Although a distinction is made here between memorability and memory strength, it is clear that these two concepts are closely related. Supposed that in a particular item group, targets are experimentally manipulated such that they are high in memory strength, it is likely that lures belonging to that same item group would also be considered to be high in memorability.

lexicality). In another line of research on the mirror effect, investigators have attempted to replicate the phenomenon with different item classes that were *experimentally* created. Typically, the manipulations used to create different item classes affected an item's strength, and hence memorability. These include (a) repetition (i.e., repeatedly present the item during study – e.g., Benjamin, 2001; Dewhurst & Anderson, 1999; Hilford, Glanzer & Kim, 1997; Morrell, Gaitan, & Wixted, 2004; Shiffrin, Huber, & Marinelli, 1995; Stretch & Wixted, 1998), (b) study duration (i.e., items were presented at different durations during study – e.g., Arndt & Hirshman, 1998; Criss & Shiffrin, 2004; Hirshman, 1995; Hirshman & Hostetter, 2000; Hirshman & Palij, 1990; Malmberg & Nelson, 2003, Ruiz, Soler, & Dasi, 2004), and (c) list length or category size (i.e., study items form categories of differing sizes – e.g., Cary & Reder, 2003; Dewhurst & Anderson, 1999; Malmberg & Murnane, 2002; Shiffrin et al., 1995). Regardless of the type of experimental manipulation, the theoretical motivation underlying these experiments was to determine whether an increase in hit rates (due to experimental increase in the item's memory strength) would also bring about a corresponding decrease in FA rates – i.e., a mirror effect.

5.4 Between-List and Within-List Strength Manipulations (Stretch & Wixted, 1998)

The recurring finding from these studies mentioned above has been that mirror effects could be produced when the strength manipulation is imposed *between lists*, but not *within list*. In "between-lists" experiments, strengthened (strong) items are presented in a separate study phase (and tested in a separate recognition test) from non-strengthened (weak) items. In contrast, when strength manipulation is imposed "within list", both the study phase and test phase consist of a mixture of strong and weak items.

An example of the contrast between within-list and between-list strength manipulation could be found in Stretch and Wixted (1998). In a series of experiments, these researchers investigated strength-based effects in the commonly observed WFE. In one experiment, the between-list strength manipulation was implemented by presenting items (low and high frequency words) *five times* each in one study phase, and only *once* in another study phase. Participants were administered a recognition

test immediately after each study phase. In this paradigm, two kinds of mirror effects were found, one which was frequency-based and one which was strength-based. The frequency-based mirror effect was simply the classic WFE pattern whereby the hit rate was higher and the FA rate was lower for low frequency than high frequency words. This was found regardless of the number of times items were presented in the study phase. The strength-based mirror effect refers to the way that for both low and high frequency words, hit rates increased and FA rates decreased in the condition where items were studied five times, relative to where they were studied only once.

When strength was manipulated within list, however, the strength-based mirror effect was not found. For example, in one experiment (Stretch & Wixted, 1998, Experiment 2), high frequency words and low frequency words were presented within the same study phase, but high frequency words were selectively strengthened as they were presented *five times* each throughout the study list while low frequency words were presented only *once*. For comparison, a control condition was carried out where there was no differential strengthening for high and low frequency items (both types were presented once during study). Of interest was whether item strengthening (through repetition) would increase the hit rate and reduce the FA rate of high frequency words. Across the two conditions, hit rate increased for high frequency words as a result of repetition during study. However, there was no concomitant decrease in the FA rate. Put another way, despite being the strong (strengthened) items in study phase, high frequency words still produced significantly higher FA rates than low frequency words, which belonged to the weak category.

5.5 Evidence for the Criterion-Shift Model: Mirror Effects from Between-List Strength Manipulations

The finding from Stretch and Wixted (1998), that a mirror effect was produced when strength was manipulated between lists, was consistent with the criterion-shift SD model. According to this account (see Figure 2), the distribution for old repeated items has shifted further to the right (to reflect an increase in strength), and the response criterion is adjusted accordingly by also shifting to the right (and thereby becoming more conservative). Because it is assumed that the strength manipulation would not affect the characteristic, and hence the location, of the new item

distribution (see also Shiffrin et al., 1995), this criterion shift would therefore result in a reduction of FA rates for items in the strengthened (strong) condition.

In other areas of research on recognition memory, the criterion-shift SD model has also been implicated. For example, Hirshman (1995) showed that the bias estimate, *C*, associated with weak items (which were studied for a short duration) varied depending on whether these items were studied and tested in a pure-list or a mixed-list context. Specifically, the bias estimate *C* indicated a more liberal criterion placement for weak items when the study and test phases contained weak items only (i.e., in a pure-list context), while a more conservative criterion was set when both weak and strong items (those that were studied for a longer duration) were included (i.e., in a mixed-list context). Despite the apparent shifting of the response criterion, Hirshman showed that the discrimination estimate (*d'*) of weak items remained stable across pure- and mixed-list conditions.

5.6 Evidence Against Criterion-Shift Models: The Absence of Mirror Effects from Within-List Strength Manipulations

In contrast to experiments where strength was manipulated between lists, experiments using within-list strength manipulations have, on the whole, failed to demonstrate the mirror effect. In Stretch and Wixted's (1998, Experiment 2) paradigm, where strength was manipulated within list, participants did not appear to take into the account that high frequency words in the study list were selectively strengthened (and therefore were more memorable), and adopt a more conservative response criterion in order to reduce FA rates for high frequency word lures. As a result, even though the hit rate for high frequency words increased because of the strength manipulation, there was no accompanying decrease in the FA rates for these items.

Consistent with Stretch and Wixted's (1998) findings are other investigations on the WFE, which showed that by using an encoding manipulation, the hit rate pattern of the WFE (i.e., low frequency words producing higher hit rates than high frequency words) could be reduced or eliminated whereas the FA rate pattern of the WFE (i.e., high frequency words producing higher FA rates than low frequency words) would remain intact. For example, participants in Hirshman and Arndt's

(1997) experiment studied high and low frequency words of varying levels of concreteness. When participants were instructed to encode the items in the context of a concreteness judgment, the typical hit rate difference between high and low frequency words did not eventuate. However, even though this encoding manipulation effectively eliminated the WFE in hit rates, the FA-rate component of the WFE appeared to be impervious to the manipulation; regardless of the type of encoding condition, FA rates remained significantly higher for high frequency words than for low frequency words. In another investigation, Criss and Shiffrin (2003) also showed that the WFE in hit rates was largely absent when study items were encoded in the context of a judgment pertaining to an item characteristic, such as concreteness, pleasantness or animacy. However, like Hirshman and Arndt, the difference in FA rates between high and low frequency words was maintained regardless of the type of encoding environment imposed.

Overall, these above studies (Criss & Shiffrin, 2003; Hirshman & Arndt, 1997; Stretch & Wixted, 1998) convincingly showed that FA rates are particularly resistant to change even when hit rates are susceptible to the effects of experimental manipulations. If participants are approximating an item group's level of memorability by its hit rate, and thereby adjusting their response criterion accordingly, hit rate changes should invariably translate to FA rate changes. In such paradigms where the item memorability fluctuates from item to item within a single test (i.e., when the strength manipulation is within list), criterion adjustments are necessary throughout the test phase such that a lower FA rate for lures of high memorability can be achieved. The stability of FA rates in these experiments therefore suggests that criterion shifts do not readily occur, especially when these adjustments are required on a trial-by-trial basis.

5.7 The Absence of Within-List Criterion Shifts: Implication on the Hension Effect

Based on Stretch and Wixted's (1998) findings, the parsimonious conclusion which could be made is that criterion shifts may occur between lists (or more correctly, between recognition tests), but they are unlikely to occur within list (i.e., within a single recognition test). In support of this conjecture, other researchers (using materials other than high and low frequency words) have found that when strength

was manipulated between lists through repetition, there was an effect on both hit rates (which increased) and FA rates (which decreased; e.g., Benjamin, 2001), but when the repetition manipulation was imposed within list, changes were only observed in hit rates while FA rates remained stable (e.g., Shiffrin et al., 1995).

The lack of evidence for within-list criterion shifts poses a critical problem for J. Brown et al.'s (1977) hypothesis that the FA rate for highly memorable lures is low because a more conservative response criterion is set for these items. Generalising from Stretch and Wixted's (1998) conclusion that participants do not shift their criterion on a trial-by-trial basis, the role of memorability in the production of the WFE and other mirror effects (such as that observed in the hension effect paradigm), has become questionable. In these paradigms, the mirror effect was argued to arise because a more conservative response criterion was adopted for test items high in memorability (i.e., low frequency words and natural words) than for test items low in memorability (i.e., high frequency words and regular nonwords). In another line of argument against the criterion-shift account for WFE, it has been pointed out that participants generally hold an erroneous perception of memorability levels among item classes: contrary to recognition performance, high frequency words were rated to be more memorable than low frequency words (Green & Thapar, 1995; Wixted, 1992). However, as discussed in the previous chapter, later studies (Benjamin, 2003; Guttentag & Carroll, 1998) and Experiment 5 in this thesis (see section 4.8) showed that memorability ratings aligned more faithfully to discrimination estimates when memorability was assessed during test. Thus, it may be premature to dismiss the possibility of within-list criterion shifts.

Nonetheless, evidence has yet to be found which indicates that participants adjust their response criterion in accordance to each test item's memorability within a single test. Such evidence is crucial in bolstering the argument that the lower FA rate achieved for natural words in the hension effect paradigm was driven by the high memorability of these items. Put another way, in order for a memorability-based criterion-shift model to be viable for the hension effect, one would have to show that within-list criterion shifts do occur. Experiment 6 was therefore carried out in an attempt to demonstrate mirror effects when a within-list strength manipulation (study duration) was imposed on the hension effect materials.

5.8 Experiment 6: Mirror Effects in the Hension Effect Paradigm – Study Duration Manipulated Within List

Experiment 6 shared two commonalities with Stretch and Wixted's (1998) investigation. First, like the high and low frequency words used by Stretch and Wixted, the hension effect item groups employed here differed in preexperimental item memorability. As Experiment 5 showed (see section 4.8), natural words are generally regarded to be more memorable than regular nonwords and irregular nonwords. In this way, like high and low frequency words, inter-group differences in item memorability existed preexperimentally for the hension effect items. The second feature common between Experiment 6 and Stretch and Wixted's research was that a strength manipulation was employed to experimentally manipulate item memorability. Thus, analogous to Stretch and Wixted's aims, the enquiry pertinent to Experiment 6 was whether strength-based effects (i.e., those produced by strength manipulations) could be observed in addition to the existing item-based effect (i.e., the hension effect).

5.8.1 Study Duration as a Strength Manipulation

In contrast to Stretch and Wixted (1998), who used repetition to manipulate item strength, the strength manipulation employed in Experiment 6 entailed presenting study items at different durations. The effects of study duration on recognition memory performance has been investigated by a number of researchers in the past (e.g., Arndt & Hirshman, 1998; Criss & Shiffrin, 2004; Hirshman, 1995; Hirshman & Hostetter, 2000; Hirshman & Palij, 1992; Malmberg & Nelson, 2003; Ruiz et al., 2004). However, most have primarily focused on these effects on hit rates, or have employed a within-subjects design where the effect on FA rates could not be properly examined (e.g., Malmberg & Nelson, 2003). Other experiments, however, have demonstrated duration effects on FA rates, but in these, the duration manipulation was implemented between lists. For example, Hirshman and Palij manipulated study duration by presenting all items at one duration in one study phase, and at a different duration in another. They found reliable effects of duration on hit rates, but duration effects on FA rates were only evident when there was a substantial difference between the two durations (800 ms versus 2,500 ms, rather than say, 800 ms versus 1,200 ms). In this comparison, FA rates were higher for the 800 ms

duration than for the 2,500 ms condition. Similarly, Ruiz et al. (2004) manipulated study duration between lists, but the durations being compared were longer in general (1 s, 3 s, 7 s, and 12 s). They found that with increasing study durations, there was an overall trend of hit rate increase and FA rate decrease, but the greatest change in hit and FA rates was between durations of 1 s and 3 s. Thus, it appears that duration effects would be most pronounced when there is at least a difference of two seconds between the durations being compared, and that both durations are reasonably short (no longer than 3 seconds). On these conclusions, the durations of 500 ms and 3 s were adopted in Experiment 6 for the short and the long duration condition respectively.

5.8.2 Predictions for the Strength-Based Effect (The Duration Effect)

In essence, the design of Experiment 6 could be compared to that found in Morrell et al. (2002, Experiment 1). In their experiment, Morrell et al. used repetition during study to manipulate the strength of word items from two different semantic categories – professions or locations. For half of the participants, profession items were presented multiple times and location items were presented only once, whereas this arrangement was reversed for the remaining half of the participants. In this way, for each participant, the strength manipulation was within list in that the study list consisted of both repeated (strong) and nonrepeated (weak) items. Consistent with Stretch and Wixted's (1998) results, Morrell et al. found that the hit rate was significantly higher for items from the strong category than for those from the weak category, but there was no difference between the two categories in FA rates.

In the same manner as Morrell et al. (2002), participants in Experiment 6 were divided into two groups. For both participant groups, the strength manipulation was made known before study. That is, participants were informed that certain types of items would be associated with certain study durations. As both regular and irregular nonwords were meaningless nonwords, these two item types were simply referred to as belonging to the "nonword" category. Natural words therefore belonged to the "word" category. For one group of participants, natural words were presented for 500 ms each and nonwords were presented for 3 s each, and vice versa for the other participant group. Thus, the durations of 3 s and 500 ms represented the strong and the weak conditions respectively, and this duration manipulation was implemented

within list as participants were presented with both strong and weak items. If the current study was to replicate previous findings (e.g., Morrell et al., 2002; Stretch & Wixted, 1998), strength-based (duration) effects would manifest in hit rates but not in FA rates. That is, within each item category (natural words, regular nonwords, irregular nonwords), the hit rate was expected to be higher in the 3 s condition than in the 500 ms condition, but FA rates were not expected to be affected by presentation duration. This pattern of results would therefore conform with previous research which failed to show evidence for within-list criterion shifts.

Although the items were treated as belonging to only two categories (nonwords and words) during the experiment, subsequent analyses would treat regular and irregular nonwords as separate item classes. In this way, strength-based effects could be scrutinised in finer detail as analyses would involve items (natural words, regular nonwords, and irregular nonwords) of three distinct levels of preexperimental memorability. Because of these pre-existing differences in item memorability, it was conceivable that duration effects might be more evident for one item type (e.g., irregular nonwords) than another (e.g., natural words). To speculate, preexperimental knowledge of natural words would enable these items to be identified and encoded in memory even if their presentation was brief. Nonwords, however, were expected to benefit substantially from a longer presentation duration. In this way, hit rate increases due to longer duration might be greater for items that were low in preexperimental memorability (i.e., nonwords). Despite this prediction of differential strength-based effects in hit rates across item classes, it was expected that if findings were to replicate those from Morrell et al. (2002) and Stretch and Wixted (1998), strength-based (duration) effects in FA rates would not emerge in any of the item categories. If, however, the FA rate of any item category varied according to the duration manipulation, this finding would indicate that correct rejections could be made on the basis of item memorability. In turn, supporting evidence would be found for the within-list criterion shift model.

5.8.3 Predictions for the Item-Based Effect (The Hension Effect)

Apart from the effect arising from the experimental strength manipulation, the hension effect, as in previous experiments reported so far in this thesis, was again of interest. Because memorability already differed preexperimentally for the hension

effect items, differences in hit and FA rates found among these item groups would also be referred to as item-based effects. In particular, because natural words are more memorable than regular nonwords, it was predicted that the mirror pattern would be produced for these two item groups. Specifically, regardless of the duration condition, the hit rate was expected to be higher and the FA rate lower for natural words than regular nonwords.

For regular and irregular nonwords, however, the predictions for Experiment 6 deviated from those made in previous experiments. It was hypothesised that the FA rate difference between regular and irregular nonwords that was typically seen in the hension effect, might be reduced by removing items that were highly similar to others from stimulus pool. This modification in the materials was motivated by a recent finding from Cleary et al. (2005). As mentioned briefly in Chapter 2 (see section 2.4), Cleary et al. (2005) demonstrated that in the hension effect materials devised by Whittlesea and Williams (2000), regular nonwords possess more orthographic neighbours (real English words that differ from the item by only one letter) than do irregular nonwords. However, a more scrupulous inspection of Whittlesea and Williams's materials revealed that *inter-stimulus similarity* maybe a critical artefact arising from the construction of nonwords with a high orthographic neighbourhood size. For example, one of Whittlesea and Williams's regular nonwords (HENSION) resembles very closely to one of the natural words (TENSION). Moreover, inter-stimulus similarity is also apparent for item pairs such as SONDER and CONDER, BINICAL and CLINICAL, and others. To obtain an index of inter-stimulus similarity for each item in the original hension effect materials, Cleary et al. counted the number of items in the stimulus pool that share the first and/or the last three phonemes with that item in question. A subsequent analysis showed that on average, regular nonwords have greater inter-stimulus similarity than both natural words and irregular nonwords. In a follow-up experiment reported in their article, Cleary et al. constructed a new set of hension effect materials, with the regular and irregular nonword groups differing in orthographic neighbourhood size, but with the two item groups being equivalent in terms of inter-stimulus similarity. In this experiment, Cleary et al. successfully reversed the regular nonword-irregular nonword FA rate difference typically seen in the hension effect – regular nonwords were found to produced significantly fewer false alarms than did irregular nonwords.

As in Cleary et al. (2005), the issue of inter-stimulus similarity was raised in the design of Experiment 6. However, rather than constructing a completely new set of hension effect materials, as Cleary et al. did, it was apparent that inter-stimulus similarity in the original hension effect materials could be reduced by discarding items which were phonologically and orthographically similar to other items in the stimulus pool. Thus, for example, the regular nonword HENSION was retained whereas the natural word TENSION was removed. Likewise, SONDER was kept but CONDER was discarded. Furthermore, from this examination of the stimulus pool, it also became evident that the three item groups differed from each other in terms of item length. On average, irregular nonwords contained significantly more letters (M = 7.40) than did natural words (M = 6.82), which in turn were significantly longer than were regular nonwords (M = 6.28, see Appendix E for details on the statistical analyses). In view of this, a number of longer items, such as the natural word FINANCIAL, and the majority of irregular nonwords exceeding seven letters in length (e.g., CRINBREELP, NERBIPAT) were removed from the stimulus pool.

In total, 60 items (20 from each item category) that were high in length or inter-stimulus similarity were removed. Thus, 120 items (40 in each item category) remained (see Appendix F for the list). Following Cleary et al., a measure of inter-stimulus similarity, in terms of the number of stimuli from the pool with which an item shares its first three or last three phonemes, was calculated for each item (that are on the final list of 120) before and after the removal of the 60 items. The removal of these items proved to reduce the overall inter-stimulus similarity of remaining materials. The index for the 120 items was significantly higher before (M = 1.26) than after (M = .68) the removal of the 60 items. Although on average, the similarity index for regular nonwords (M = 1.58) was still higher than that for irregular nonwords (M = .33), this main effect was qualified by an interaction showing that the decrease in the similarity index, as a result of the removal of "similar" items from the stimulus pool, was significant for regular nonwords, but not for irregular nonwords (see Appendix G for details). Details regarding bigram frequency measures and item length of the remaining 120 items can also be found in Appendix D and E respectively.

In view of the decrease in inter-stimulus similarity, particularly for regular nonwords, it was expected that the FA rate difference between regular and irregular nonwords, previously observed to be significant in Experiments 1 – 5, might be reduced in Experiment 6. However, even when inter-stimulus similarity was controlled, Cleary et al. still showed that a significantly lower FA rate for natural words than regular nonwords. Thus, that component of the hension effect was not expected to be eliminated with these edited materials. That is, the mirror effect was predicted for the recognition performance between natural words and regular nonwords.

5.8.4 Summary of Predictions for Experiment 6

In summary, two sets of predictions were made for Experiment 6. First, the experimentally-imposed study duration manipulation was hypothesised to produce hit rate effects. Because this strength manipulation was applied within list, duration effects on FA rates might not eventuate as past research (e.g., Morrell et al., 2002; Stretch & Wixted, 1998) suggested that memorability-based rejections, expressed as within-list criterion shifts in SD models, do not take place in such paradigms. It was also speculated that strength-based effects might vary across the three groups of hension effects items, because the three item groups differ in terms of preexperimental memorability. The second set of predictions related to the consequence of reducing the inter-stimulus similarity within the materials used in the hension effect paradigm. Following Cleary et al's (2005) results, it was hypothesised that the FA rate difference between regular and irregular nonwords (one part of the hension effect) might diminish in Experiment 6 as items contributing to high inter-item similarity were removed from the stimulus pool.

5.8.5 Method

Participants. The participants were 48 students from the University of Southampton. These students were either psychology undergraduates who participated in return for course credits, or non-psychology students who were reimbursed £4 for their time. All spoke English as their only fluent language and none has participated in the experiments reported in previous chapters.

Materials and design. The materials consisted of items from the three categories (natural words, regular nonwords, irregular nonwords), as adopted from Whittlesea and Williams (2000). As detailed above (see section 5.8.3), the original set of 60 items in each category was trimmed down to 40 items per category such that similarity among items could be reduced. The edited set of materials is listed in Appendix F.

The 48 participants in Experiment 6 were divided into two groups of 24. For one participant group (hereafter the "long-nonwords" group), natural words were presented for 500 ms and nonwords (regular and irregular) 3 s each during study. For the other participant group (hereafter the "long-words" group), natural words were presented for 3 s and nonwords (regular and irregular) 500 ms each during study. Counterbalancing ensured that across participants, each word and nonword item served equally often as old and new.

Procedure. Participants were tested in a quiet room in groups of up to four. The room was partitioned such that each participant was seated in front of a Macintosh computer. The standard set of study phase instructions were given to participants (see Appendix B), where they were also informed of the presentation duration of the study items. Participants in the long-nonwords group were told that English words would be presented for half a second and non-English words for 3 seconds, whereas the durations were reversed in the instructions for the long-words group.

The study phase consisted of 60 study items (20 from each of the three item categories), and the three additional practice items as used in Experiment 1. Likewise, the manner in which study items were presented was identical to that in previous experiments (see section 2.2.1 for more details). Regardless of the item's presentation duration (500 ms or 3 s), the ITI was 1 s.

After the study phase, participants were given a 10 minute retention interval[12], during which they played a computer game. Following this interval, the test phase was administered. Participants were first given the standard set of test instructions, where they were reminded of the types of items (words and nonwords) that were studied earlier, and that the two item types were presented at different durations. Participants were told to make their recognition judgments by pressing the right key ("p" on the keyboard) for "old" and the left key ("q" on the keyboard) for "new". The same six practice items from previous experiments were again used here at the start of the test phase. The test phase proper consisted of 120 items (40 items from each of the 3 categories). Throughout the test phase, participants were reminded of their task requirement by the instruction label ("Left for NEW or Right for OLD"), which was placed near the top on the screen. Apart from these procedural details, the manner in which items in the test phase were presented was largely identical to that in previous experiments. The only exception was that unlike previous experiments (Experiments 1 – 4), where participants were allowed a break opportunity at the half-way point in the test phase (after 90 test items), the 120 test items here were presented in a singular block (i.e., no break opportunity was offered).

5.8.6 Results

Mean hit rates and FA rates obtained in Experiment 6 for the three item types, in short and long duration conditions, are shown in Table 14. Different typefaces (italicised verses non-italicised) are used to specify the two participant groups. As described earlier, the long-nonwords group was presented with words at a short duration and nonwords at a long duration, and this was reversed for the long-words group. The same analysis – a 3 (item type: natural word/ regular nonword/ irregular nonword) x 2 (study duration: short/long) mixed ANOVA, with item type as the

[12] Because there are 20 word and 40 nonword study items, the duration of the entire study phase for the long-word and the long-nonword groups are unequal. Taking ITI into account, the study phase would take 140 s for the long-word group and 190 s for the long-nonword group. Thus, the study phase is 50 s shorter for one group than the other. It could be argued that this difference would become insignificant in the context of the 10 minute retention interval. The inclusion of a retention interval between study and test is a common feature in recognition memory experiments, and on the basis of past research, should have no unforeseen effects on the predictions.

within-group factor and study duration as the between-group factor – was conducted on each of the three measures: hit rate, FA rate and the discrimination estimate, d' [13].

Hit Rate. The analysis on hit rates showed that there was an item main effect, $F(2, 92) = 29.58, p < .001, MSE = .018, \eta^2 = .391$. Post-hoc paired-samples t tests ($\alpha = .0167$) showed that averaged across study durations, natural words ($M = .78$) produced a higher hit rate than did regular nonwords ($M = .64$), $t(47) = 4.75, p < .001$, SE = .030, $\eta^2 = .324$, which in turn produced a higher hit rate than did irregular nonwords ($M = .57$), $t(47) = 2.73, p < .01, SE = .024, \eta^2 = .137$. Additionally, there was a main effect of study duration, $F(1, 46) = 5.45, p < .05, MSE = .012, \eta^2 = .106$. Averaged across item type, the hit rate was higher when items were studied at a long duration ($M = .70$) than when they were studied at a short duration ($M = .63$). This main effect of duration was qualified by a significant item x encoding interaction, $F(2, 92) = 4.52, p < .02, MSE = .018, \eta^2 = .089$. Post-hoc independent-samples t tests ($\alpha = .0167$) showed that this interaction was driven by a significant difference in the hit rates between short and long study duration for irregular nonword items only, $t(46) = 3.28, p < .01, SE = .050, \eta^2 = .067$. The corresponding comparison was not significant for natural words, $t(46) = .053, p > .90$, or for regular nonwords, $t(46) = 1.37, p > .15$.

FA Rate. As in the hit rate data, there was also a reliable main effect of item in the FA rate data, $F(2, 92) = 4.12, p < .02, MSE = .019, \eta^2 = .082$. Post-hoc t tests ($\alpha = .0167$) showed that only the comparison of FA rates between natural words ($M = .21$) and regular nonwords ($M = .29$) was significant, $t(47) = 2.59, p < .0167, SE = .031, \eta^2 = .125$. The FA rates of regular and irregular nonwords were not significantly different from each other, $t(47) = 1.74, p > .08$. Unlike the analyses on the hit rate data, however, there was no significant main effect of encoding condition, and no significant interaction (both Fs < 1).

---

[13] Some readers might note that analyses on the bias estimate, C, are not performed here. This might seem puzzling because the bias estimate, in theory, reflects the placement of the criterion, and should be particularly pertinent to the discussion of within-list criterion shifts. However, as will be addressed later in Chapter 7, recent arguments (e.g., Roediger & McDermott, 1999; Wickens & Hirshman, 2000) have highlighted the problems of interpreting C in terms of criterion shifts (see section 7.1 for a detailed discussion). On these grounds, it was therefore decided that analyses on C would not be reported.

Table 14. Experiment 6: Mean hit and FA rates for natural words, regular nonwords and irregular nonwords in the two duration conditions (short and long). Data from the long-nonwords group (n = 24) are italicised, and from the long-words group (n = 24), non-italicised. Standard deviations are in parentheses.

	Short		Long	
	Hit	FA	Hit	FA
Natural	*.78 (.11)*	*.23 (.17)*	.78 (.16)	.19 (.14)
Regular	.61 (.18)	.30 (.16)	*.67 (.13)*	*.28 (.15)*
Irregular	.49 (.16)	.26 (.17)	*.65 (.18)*	*.24 (.12)*

Discrimination Estimate. The means of estimates of discrimination (d') are presented in Table 15. Analyses on d' revealed a significant main effect of item only, $F(2, 92) = 20.47$, $MSE = .506$, $\eta^2 = .308$. Post-hoc comparisons ($\alpha = .0167$) showed that discrimination was significantly better for natural words ($M = 1.84$) than for both regular nonwords ($M = 1.06$), $t(47) = 4.54$, $p < .001$, $SE = .172$, $\eta^2 = .305$ and irregular nonwords ($M = 1.01$), $t(47) = 4.90$, $p < .001$, $SE = .168$, $\eta^2 = .338$. Estimates of d' for regular nonwords and irregular nonwords did not differ significantly from each other, $t(47) = .567$, $p > .55$.

Table 15. Experiment 6: Mean estimates of d' for the three item types in the two duration conditions (short and long). Italicised and non-italicised data correspond to the long-nonwords group (n = 24) and the long-words group (n = 24) respectively. Standard deviations are in parentheses.

	Short		Long	
Natural	*1.79*	*(.75)*	1.89	(.89)
Regular	.88	(.38)	*1.24*	*(.78)*
Irregular	.70	(.48)	*1.33*	*(.79)*

5.8.7 Discussion

The primary aim in Experiment 6 was to acquire evidence that when an item's memorability was experimentally enhanced through a strength manipulation, participants would be able to use this information to guide their recognition decisions for lures on a trial-by-trial basis. Investigations have been carried out in the past expressly on this purpose (e.g., Morrell et al. 2002; Stretch & Wixted, 1998). However, unlike these previous investigations, the present experiment utilised study duration, rather than repetition, to manipulate strength. Despite this methodological change, the pattern of results obtained from Experiment 6 paralleled those earlier studies. When strength was manipulated within list, such as presenting a proportion of study items for a longer duration, strength-based effects on recognition performance were only evident in hit rates, but not in FA rates. This was the case even when it was explicit that a certain category of the items (words or nonwords) were differentially strengthened during study. In Experiment 6, the overall hit rate (averaged across item types) increased significantly as a result of longer study durations. This improvement in hit rates was not accompanied by a concomitant decrease in FA rates. It appeared that although there was objective evidence of enhanced strength, and in turn, memorability for items studied at longer durations (as evinced by the increase in hit rates), participants were unable to incorporate memorability information in making recognition judgments, such that more new items from the memorable category could be correctly rejected.

Although Experiment 6 principally replicated previous findings (e.g., Morrell et al., 2002; Stretch & Wixted, 1998), a novel result has emerged from the experiment. As hypothesised, strength-based effects on hit rates appeared to be shaped by memorability levels that were intrinsic to item characteristics. The significant item by duration interaction indicated that hit rate increase was statistically reliable only for irregular nonwords. Increase in hit rates was observable (but not statistically significant) for regular nonwords whereas it was virtually absent for natural words. These findings were not wholly unexpected when considering that the three item groups differed in terms of their preexperimental levels of item memorability. Natural words are meaningful stimuli encountered by participants in everyday language and therefore could be easily identified and encoded in memory

even when presentation duration was short. To an extent, because regular nonwords consist of real-word components, their wordlikeness might also afford these items some level of memorability, and hence ease of identification and encoding even when presented at short durations. In contrast, irregular nonwords consist of unusual letter combinations which not only rendered these items to be unmemorable, but also created difficulty in their identification and encoding. It was therefore found that the short study duration condition greatly impaired participants' ability to recognise these items at test. Put another way, however, one could interpret that irregular nonwords reaped more benefits from longer study durations than natural words and regular nonwords, and this was reflected in the way a strength-based effect on hit rates was found for irregular nonwords but not for the other two item groups.

With the finding of category-specific effects on hit rates in Experiment 6, one potential concern was that the absence of duration effects on FA rates was due to the stable hit rates found for two of the three item types – natural words and regular nonwords. Based on this reasoning, duration effects on FA rates might be found in irregular nonwords only, for which there was a clear duration effect on the hit rate. This speculation was not supported, however, as the FA rates for irregular nonwords in the two duration conditions were close to identical (.26 for short, .24 for long). Thus, even though the memorability of irregular nonwords was enhanced in the long duration condition (as evinced by the hit rates), there was no evidence that this increase in memorability was used as a basis to correctly reject irregular nonword lures.

Apart from these strength-based effects arising from the duration manipulation, item-based effects on hit and FA rates were also expected. The hit rate pattern among the three item types conformed with the findings from previous experiments reported in this thesis. Specifically, natural words generally produced more hits than did regular nonwords. This result contradicted Whittlesea and Williams's (e.g., 1998, 2000) argument that response bias might be more lenient for regular nonwords than for other items, such that "old" judgments were more prevalent for regular nonwords, leading to both a higher FA rate and a higher hit rate for these items than for other item groups. The exact reason which led to different results being obtained here is unclear, although it had been speculated in the previous chapter (see

section 4.3) that Whittlesea and Williams's results might partly be due to the self-paced procedure they had adopted for their study phase, which enhanced the role of recollection plays in the recognition of regular nonwords. With study durations controlled (as in experiments conducted in this thesis), the involvement of recollection processes would be significantly higher for natural words than for regular nonwords, thus resulting in the hit rate advantage observed here for words. As will be apparent in the coming chapters, this outcome was replicated in the remaining experiments conducted for this thesis. Because this finding has been discussed at length here and in previous sections, it will not be addressed further in future sections.

An additional hypothesis made for Experiment 6 was that by decreasing inter-item similarity, the FA rate difference between regular and irregular nonwords might also be diminished (Cleary et al., 2005). The results obtained here conformed with this prediction. In contrast to previous experiments reported in this thesis, Experiment 6 showed that the FA rate for regular nonwords was not significantly higher than that for irregular nonwords. This outcome appeared to arise purely through the removal of a portion of items from the original stimulus set (as devised by Whittlesea & Williams, 2000), such that the degree of similarity among items was lowered. It is clear that the process of revising the experimental materials would not have any impact on the sense of discrepancy perceived for items at test. As such, the elimination of the hension effect between regular and irregular nonwords in Experiment 6 cannot be explained by the discrepancy-attribution hypothesis (e.g., Whittlesea & Williams, 1998, 2000). Thus, in line with the conclusions drawn in earlier chapters of this thesis (see Experiments 1 – 4, Chapters 1 – 3), it is increasingly doubtful that the discrepancy-attribution hypothesis is a fitting account for the hension effect.

5.9 Concluding Remarks for Chapter 5

Despite the partial elimination of the hension effect (i.e., for regular and irregular nonwords), the mirror pattern (and hence the FA rate difference) was again observed between regular and natural words in Experiment 6. As an alternative to the discrepancy-attribution account, it was argued that the memorability of natural words allowed these items to be correctly rejected more frequently than the less memorable regular nonwords. However, to support this argument, evidence is required which

shows that the memorability is assessed and incorporated into the recognition decision process on an item-by-item basis during test. In SD terms, this use of memorability in recognition judgments is manifested as within-list criterion shifts, as participants are argued to adopt a more conservative response criterion for memorable than for unmemorable items. In turn, within-list criterion shifts would lead to a suppression of FA rates for memorable items, relative to unmemorable items. Thus far, no evidence has been found which supports this memorability-based hypothesis, either from previous research (e.g., Morrell et al., 2002; Stretch & Wixted, 1998) or from Experiment 6. This issue will be addressed through further experimentation in the next chapter.

Chapter 6

6.1 Using Colour to Cue Item Memorability (Stretch & Wixted, 1998)

The overall stability of FA rates found in Experiment 6 suggests that on an item-by-item basis, memorability has little influence on recognition judgments and thus, in SD terms, participants are reluctant to adjust their response criterion repeatedly within a recognition test. Although Stretch and Wixted (1998) speculated that within-list criterion shifts *could* occur, there has been little experimental evidence to support this case. In other experiments, Stretch and Wixted showed that even when the within-list strengthening manipulation was made salient by colour cues and explicit instructions (to encourage participants to consider item memorability in their recognition judgments), FA rates remained stable despite clear strength-based effects in hits. For example, in one experiment (see Stretch & Wixted, 1998, Experiment 5), half of the items in each item category (high and low frequency words) were strengthened through repetition during study. At test, these strengthened (strong) targets were presented in one colour (e.g., red) while nonstrengthened (weak) targets were presented in another colour (e.g., green). For the lures, half in each category were arbitrarily presented in red while the remaining half were presented in green. Before test, participants were made aware of the systematic way by which test items would be coloured. In this experimental design, it was assumed that participants would be able to use the test item's colour as an indicator of its memorability level. Hence, participants would demand convincing memorial evidence (i.e., set a conservative response criterion) for red items which came from the strengthened (strong) category. For green items from the nonstrengthened (weak) category, such compelling evidence was not required, and therefore a liberal criterion would be set. Consequently, the FA rate was expected to be lower for red than for green lures. Results, however, showed that the FA rates were not different between red and green lures.

6.2 Problems with Stretch and Wixted's (1998) Colour Cueing Paradigm

The absence of strength-based effects in FA rates led Stretch and Wixted (1998) to conclude that participants were not able to shift their response criterion in accordance to the memorability of each individual test item. However, this conclusion

would only be valid if the colour cues and instructions were fully and correctly utilised by participants. That is, the lack of evidence for memorability-based FA suppression might be due to the way that the colour-cueing system employed by Stretch and Wixted was not as clear as was intended. Specifically, a potential ambiguity might arise if participants did not assume that items studied repeatedly will be presented in red only, and never in green during test, and items studied once will be presented in green only, and never in red during test. Consider the scenario where a participant was presented with a new item in red, and the decision to be made was simply whether it was studied before (old) or not (new). Although it was specified that item presented in red would have been studied *multiple* times earlier, if the participant was under the suspicion that the item was actually only studied *once* (not multiple times as the colour indicated), he/she would still have to judge the item as old, rather than new. Thus, distrust of the veracity of the colour cues might preclude participants from consistently using item colour to perform memorability-based correct rejections. Likewise, although Morrell et al. (2002) informed their participants that one of the item categories (profession or location) would be differentially strengthened during study, it was still conceivable that participants might mistakenly believe that a new item belonging to the strengthened (strong) category was in fact studied only once, and therefore it should receive an "old" response.

A similar line of argument could also be applied to the lack of strength-based effects on FA rates in Experiment 6, as reported in the previous chapter. In Experiment 6, it was assumed that participants were able to infer memorability from the lexical status of the item. That is, participants in the "long-words" group, and those in the "long-nonwords" group, would regard item memorability to be enhanced (through the long study duration) for words and nonwords respectively. However, to the extent that participants were erroneous in their lexical judgments, the lexicality of the items would become an unreliable cue for memorability. Relevant to this speculation is the findings from the LDT, as detailed earlier in Chapter 2 (see section 2.4.2) – participants produced errors in lexical decision for a sizeable portion of items, particularly in the case of regular nonwords. Thus, it was plausible that participants mistook a nonword for a word, or vice versa; and subsequently came to believe that item lexicality was an imperfect cue for memorability.

With these considerations in mind, it was therefore important to discount the possibility that the ineffectiveness of the cues was precluding participants from using memorability as a basis for correct rejections. In other words, evidence for within-list criterion shifts could still be attained if conditions were made optimal for memorability-based information to be incorporated in recognition judgments. This hypothesis was tested in Experiment 7, in which Stretch and Wixted's (1998, Experiment 5) paradigm was slightly modified by the inclusion of a clear test decision label throughout the course of the test phase.

6.3 Experiment 7: The Use of Effective Memorability Cues in Producing Strength-Based Mirror Effects (I)

To circumvent any possible confusion regarding colour as a veritable indicator of item strength or memorability, each test item in Experiment 7 was accompanied by a label which would specify the decision participants were required to make for that particular item. Similar to the procedure used by Stretch and Wixted (1998, Experiment 5), half of the items from each category in the hension effect materials were studied multiple times (three times) at test and the remaining items were studied once. At test, targets that were studied three times were presented in red, whereas those studied only once were presented in blue. For new items, half were presented in red and half in blue. However, unlike Stretch and Wixted's procedure, each test item was accompanied by labels which clarified the dichotomous decisions to be made – "studied three times or new" for red items, and "studied once or new" for blue items. These labels made it clear to participants that colour was a reliable indicator of the item's expected strength or memorability if it had been studied, and based on this information, participants could then reject the highly memorable red lures more readily than the less memorable blue lures.

6.3.1 Method

Participants. Twenty-four students from University of Southampton took part in this experiment in return for course credit. For all participants, English was the only language they could speak fluently, and none had participated in previous experiments reported in this thesis.

Materials and design. The materials used here were identical to those in Experiment 6 – i.e., 40 items from each of the three categories (natural words, regular nonwords and irregular nonwords). These 120 items (in total) were selected from the original set of stimuli provided by Whittlesea and Williams (2000; see section 5.8.3 for selection details and Appendix F for the listing of items used in this experiment).

For the 40 items from each category, study status (old or new) was crossed with repetition ("studied once" or "studied three times") to create four counterbalancing conditions, and items were assigned such that across participants, they appeared equally often in each counterbalancing condition. At study, half (20) of the items from each category were presented. Of these 20, 10 were studied three times and 10 were studied once. Hence, together there were 60 study items – 30 were studied three times and 30 were studied once. All 120 items (40 from each category) were presented in the test phase. Of these, 60 were old and 60 were new. At test, repetition was designated by colour such that targets studied three times were presented in red and targets studied once were presented in blue. For the lures (20 from each category), half (10) were presented in red and half in blue.

Procedure. The procedure for the study phase in Experiment 7 was identical to that in Experiment 6, apart from the following exceptions. One group of participants were tested. They were informed that during the study phase, half of the items would be presented once, and the other half would be presented three times. With half of the items being repeated, the study phase proper consisted of 120 trials. Each study item was presented for a fixed duration of 1 s, with an ITI of 1 s.

The test phase followed immediately after study. Here, participants were instructed that they were to make an old/new recognition decision for each test item. Further, they were told that should studied items be presented in the recognition test, those that were studied three times would be shown in red, and those studied once would be shown in blue. They were also told that for items not studied earlier, half would be shown in red and half in blue. At test, if the item was presented in blue, the label "Press LEFT key if NEW" was presented to the left of the test item, and the label "Press RIGHT key if STUDIED ONCE" was presented to the right of the test item. If the test item was presented in red, the labels were "Press LEFT key if NEW" and "Press RIGHT key if STUDIED 3 TIMES" to the left and the right of the item

respectively. Apart from these procedural details, the methodology of Experiment 7 was the same as that described in Experiment 6 (see section 5.8.5 for more details).

6.3.2 Results

As can be seen in Table 16, for each item category (natural words, regular nonwords, irregular nonwords), there is a hit rate for items studied once (blue targets), a hit rate for items studied three times (red targets), a FA rate for new blue items (blue lures), and a FA rate for new red items (red lures). Additionally (see Table 17), for each item category, an estimate of *d'* (discrimination) was calculated separately for items presented in blue, and for items presented in red.

Hit Rate. A 3 (item: natural words/ regular nonwords/ irregular nonwords) x 2 (repetition: once[blue]/three times[red]) repeated-measures ANOVA was performed on the hit rate data. The analysis revealed a significant main effect of item, $F(2, 46) = 7.42$, $p < .01$, $MSE = .021$, $\eta^2 = .244$, and post-hoc paired-samples t tests ($\alpha = .0167$) showed that averaged across the repetition conditions, the natural word hit rate ($M = .82$) was significantly higher than the irregular nonword hit rate ($M = .70$), $t(23) = 3.81$, $p < .001$, $SE = .030$, $\eta^2 = .387$. The hit rate comparison between regular nonwords ($M = .75$) and natural words, and that between regular nonwords and irregular nonwords, were both marginally significant, $t(23) = 1.83$, $p < .08$, and $t(23) = 2.18$, $p < .04$, respectively. The repetition main effect was also found to be significant, $F(1, 23) = 79.14$, $p < .001$, $MSE = .017$, $\eta^2 = .775$. Averaged across item types, hit rate was significantly higher for items studied three times ($M = .85$) than for items studied only once ($M = .66$). There was no evidence of an item x repetition interaction, $F(2, 46) = 1.80$, $p > .15$.

FA Rate. The same 3 (item) x 2 (repetition) analysis performed on hit rates, was also performed on FA rate data. This analysis revealed a significant main effect of item, $F(2, 46) = 5.65$, $p < .01$, $MSE = .024$, $\eta^2 = .197$. Collapsed across repetition/colour, post-hoc paired-samples t tests ($\alpha = .0167$) showed that significantly fewer false alarms were produced for natural words ($M = .11$) than for regular nonwords ($M = .20$) and irregular nonwords ($M = .20$), $t(23) = 2.95$, $p < .01$, $SE = .030$, $\eta^2 = .275$, and $t(23) = 2.67$, $p < .0167$, $SE = .036$, $\eta^2 = .237$, respectively. FA

rates for regular and irregular nonwords did not differ significantly from each other, $t(23) = .215, p > .80$.

Table 16. Experiment 7: Mean hit and FA rates (standard deviations are in parentheses) for the three item groups, presented in blue and in red ($N = 24$).

	Blue ("studied once")		Red ("studied three times")	
	Hit	FA	Hit	FA
Natural	.74 (.18)	.12 (.14)	.89 (.14)	.10 (.12)
Regular	.66 (.17)	.23 (.18)	.85 (.15)	.16 (.13)
Irregular	.58 (.19)	.27 (.18)	.82 (.14)	.14 (.12)

Crucially, there was also a significant main effect of repetition on FA rates, $F(1, 23) = 7.56, p < .02, MSE = .026, \eta^2 = .247$. That is, averaged across item types, new items produced significantly lower FA rates if they were presented in red (M = .13) rather than in blue ($M = .21$). The item x repetition interaction indicated a trend towards significance, $F(2, 46) = 2.53, p < .10, MSE = .012, \eta^2 = .099$. A cursory examination of the data suggests that the effect of repetition (lower FA rates for red lures than blue lures) was more evident in irregular nonwords (difference in FA rates for red and blue items = .13) than regular nonwords (difference = .07), which in turn showed, numerically, a greater repetition effect than natural words (difference = .03). Paired-samples t tests ($\alpha = .0167$) supported this observation in that the repetition effect on FA rates was statistically significant for irregular nonwords, $t(23) = 3.12, SE = .040, p < .01, \eta^2 = .298$, marginally significant for regular nonwords, $t(23) = 1.87, p < .08, SE = .038, \eta^2 = .132$, and nonsignificant for natural words, $t(23) = .76, p > .45$.

Discrimination Estimate. The analysis on d' – again a 3 (item) x 2 (repetition) repeated-measures ANOVA – showed a significant main effect of item, $F(2, 46) = 18.75, p < .001, MSE = .331, \eta^2 = .449$. The item main effect arose because averaged across repetition, discrimination for natural words ($M = 2.22$) was significantly better than that for regular nonwords ($M = 1.70$) and irregular nonwords ($M = 1.53$), $t(23) = 4.20, p < .001, SE = .124, \eta^2 = .434$, and $t(23) = 6.74, p < .001, SE = .146, \eta^2 = .664$,

respectively ($\alpha = .0167$). Discrimination estimates for regular and irregular nonwords did not differ significantly from each other, $t(23) = 1.36$, $p > .15$. Collapsed across item types, discrimination was significantly better for items presented in red ($M = 2.26$) than in blue ($M = 1.37$), $F(1, 23) = 120.20$, $p < .001$, $MSE = .236$, $\eta^2 = .839$. In addition, there was also an item x repetition interaction which approached significance, $F(2, 46) = 3.15$, $p < .06$, $MSE = .261$, $\eta^2 = .121$. Although for all three item types, discrimination improved as a result of repetition, the improvement was most pronounced for irregular nonwords (difference in *d'* between red and blue items = 1.16), and was intermediate for regular nonwords (difference = .86), and was least pronounced for natural words (difference = .64).

Table 17. Experiment 7: Mean estimates of *d'* (standard deviations in parentheses) for the three items types, presented in blue and in red ($N = 24$).

	Blue ("studied once")	Red ("studied three times")
Natural	1.90 (.77)	2.54 (.57)
Regular	1.27 (.81)	2.13 (.61)
Irregular	.95 (.65)	2.11 (.56)

6.3.3 Discussion

As in previous research (e.g., Morrell et al., 2002; Shiffrin et al., 1995; Stretch & Wixted, 1998) and in Experiment 6 (see section 5.8), the results from the current experiment showed that strong (strengthened) items produced a higher hit rate than weak (nonstrengthened) items. In SD terms, this finding could simply be explained by assuming separate distributions for strong and weak targets (see Figure 2, section 5.2). On the underlying continuum, the distribution of strong targets is located to the right of the distribution of weak targets. Unless the displacement of the response criterion (to the right) is enough to compensate for the distribution shift for the strong targets, there would be an increase in hit rates as a consequence of the strength manipulation.

For FA rates, it was notable that the item-based effect (i.e., the hension effect) was again only observed between regular nonwords and natural words, but not

between regular and irregular nonwords. This finding was identical to the result in Experiment 6, where the same set of materials, with reduced levels of inter-stimulus similarity, was used. The replication here lends further support for the argument that the FA rate difference between regular and irregular nonwords is not likely to be accounted for by the discrepancy-attribution hypothesis (e.g., Whittlesea & Williams, 1998, 2000). Instead, it appears that the elevated FA rate of regular nonwords, as observed in the hension effect, might be partially caused by the high inter-stimulus similarity within that item group.

The unique finding from Experiment 7, however, was that FA rates for the lures belonging to the strong item group (red) was significantly lower than FA rates for lures belonging to the weak item group (blue). Such finding contradicts those found in previous research (e.g., Morrell et al., 2002; Stretch & Wixted, 1998), and is consistent with SD accounts which permit within-list criterion shifts (e.g., J. Brown et al., 1977). It is notable that apart from the materials and the test labels used, this current experiment was almost identical to Stretch and Wixted's study (1998, Experiment 5) in terms of design and procedure. Thus, it could be argued here that the implementation of test labels in Experiment 7 was critical in producing the previously elusive result of within-list strength-based effects on FA rates. It was suggested that these test labels alleviated the potential confusion participants might experience in regards to the reliability of colour as a cue to item memorability. The labels stipulated that lures from the strong class (presented in red) were either "studied 3 times" or "new". In this way, the problematic possibility that participants might believe the red lure was in fact presented once (and therefore warranting an "old" response) was eliminated. Consequently, participants were in a better position to integrate memorability-related information into their recognition judgments, leading to a suppression of FA rates for lures from the strong category, relative to lures from the weak category.

However, an alternative explanation for the presence of a strength-based effect in FA rates here, but not in Stretch and Wixted's (1998) experiment, was that the materials used in the present experiment consisted of both words and nonwords, whereas Stretch and Wixted used words only. Indeed, there was some indication from the present experiment that memorability was a factor in rejecting new items only

when the items were nonwords, rather than natural words. The analysis on FA rates showed that the item by repetition interaction approached significance (see also Table 16) – FA reduction for red items (relative to blue items) was most pronounced in irregular nonwords, followed by regular nonwords, and least pronounced in natural words. Intriguingly, it appears that the involvement of memorability in the correct rejections of lures might be inversely proportional to the wordlikeness of the test item.

Perhaps this pattern of FA rates could be explained by invoking the notion of an *intrinsic* level of memorability that is associated with each item type (Ghetti, 2003). As reported earlier in Experiment 5 (see section 4.8), natural words were rated to be more memorable than regular nonwords, which were in turn rated to be more memorable than irregular nonwords. It seems that preexperimental experience associated with the item was used as a direct benchmark against which memorability was rated. Furthermore, natural words possess meaning, which in turn would aid their encoding during study. Meaningfulness, therefore, may allow natural words to be perceived as more memorable than nonwords. In contrast, experimentally-imposed manipulations (e.g., repetition of study episodes, length of study duration) could be said to be directed at the *extrinsic* memorability of the item (Ghetti, 2003). It follows then that effects arising from manipulating items' extrinsic memorability might not manifest as clearly in stimuli that are already intrinsically memorable. That is, if natural words were deemed to be highly memorable, memorability would be used to reject new natural words even when they belonged to the weak item category (i.e., presented in blue at test). This hypothesis could be supported by the generally low FA rates observed for both red and blue new natural words (see Table 16). On the other hand, memorability might become a more useful tool in making correct rejections for items that are not intrinsically memorable (e.g., irregular nonwords). Based on this line of reasoning, if encoding conditions were made less favourable, FA rates should increase, and memorability might come into play in correct rejections not only for the intrinsically unmemorable nonwords, but also for the intrinsically memorable natural words. The following Experiment 8 was conducted to address this hypothesis.

6.4 Experiment 8: The Use of Effective Memorability Cues in Producing Strength-Based Mirror Effects (II)

The procedure of Experiment 8 differed from that of Experiment 7 in two aspects. First, all study items in Experiment 8 were presented for 500 ms each, rather than 1 s as in Experiment 7. Second, a retention interval of 10 minutes was inserted between study phase and test phase. With these two changes, it was expected that overall, recognition performance would be impaired – hit rates would decrease and FA rates would increase in comparison to those found in Experiment 7. However, the more important purpose of these two procedural changes was to increase the *perceived difficulty* of the recognition test, as task difficulty (an extrinsic factor) would affect participants' subjective evaluation of item memorability. In Experiment 7, when study duration was longer and testing immediately followed study, natural words in general were considered as highly memorable, regardless of whether they had been studied once or three times. Under more demanding test conditions (where participants would expect their recognition performance to be poor), the number of times an item had been studied (as indicated by its colour and test label) would become a more crucial factor when its memorability is being assessed. Thus, it was expected that the effect of memorability on FA rates would also emerge for natural words in Experiment 8.

6.4.1 Method

Participants. The participants were 24 undergraduate psychology students from the University of Southampton. Course credits were given in return for participation. These participants all spoke English as their only fluent language, and none had participated in experiments reported in previous sections of this thesis.

Materials and design. The materials and design used in Experiment 8 were identical to those used in Experiment 7 (see section 6.3.1).

Procedure. The procedure in Experiment 8 was identical to that in Experiment 7, with the exception that all items in the study phase were presented for 500 ms instead of 1 s, and that at the end of the study phase, participants were given a 10 min retention interval during which they played a computer game. The test phase, with an

identical format as that in Experiment 7, was administered following the retention interval.

6.4.2 Results

As in Experiment 7, separate 3 (item: natural words/ regular nonwords/ irregular nonwords) x 2 (repetition: once [blue]/ three times [red]) repeated-measures ANOVAs were performed on the three sets of data obtained: hit rates, FA rates, and estimates of discrimination (*d'*).

Table 18. Experiment 8: Mean hit and FA rates (standard deviations in parentheses) for the three item groups, presented in blue and in red ($N = 24$).

	Blue ("studied once")		Red ("studied three times")	
	Hit	FA	Hit	FA
Natural	.72 (.20)	.21 (.15)	.85 (.14)	.13 (.12)
Regular	.58 (.19)	.37 (.18)	.78 (.16)	.26 (.16)
Irregular	.51 (.20)	.29 (.18)	.66 (.22)	.25 (.14)

Hit Rate. Overall, analyses on hit rates (see Table 18) mirrored those found in Experiment 7. There was a significant main effect of item, $F(2, 46) = 13.50$, $p < .001$, $MSE = .036$, $\eta^2 = .370$, and of repetition, $F(1, 23) = 26.30$, $p < .001$, $MSE = .035$, $\eta^2 = .533$. The item x repetition interaction was not significant, $F(2, 46) < 1$. Post-hoc analyses ($\alpha = .0167$) showed that averaged across blue and red items, natural words ($M = .79$) produced a significantly higher hit rate than did regular nonwords ($M = .68$), $t(23) = 2.77$, $p < .0167$, $SE = .038$, $\eta^2 = .250$. In turn, the hit rate was significantly higher for regular than irregular nonwords ($M = .59$), $t(23) = 2.82$, $p < .01$, $SE = .034$, $\eta^2 = .257$. The repetition main effect arose because the mean hit rate for repeated (red) items ($M = .76$) was significantly higher than that for nonrepeated (blue) items ($M = .60$).

FA Rate. The item and repetition main effects were also found to be significant in the analysis on FA rates (see Table 18), $F(2, 46) = 8.03$, $p < .002$, *MSE*

= .033, η^2 = .259, and $F(1, 23) = 12.60$, $p < .002$, $MSE = .019$, $\eta^2 = .354$, respectively. Post-hoc analyses ($\alpha = .0167$) showed that regular nonwords ($M = .31$) produced significantly more false alarms than did natural words ($M = .17$), $t(23) = 3.93$, $p < .001$, $SE = .037$, $\eta^2 = .402$. The difference in FA rates between irregular nonwords ($M = .27$) and natural words was marginally significant, $t(23) = 2.31$, $p < .03$, but the comparison between irregular and regular nonwords was not significant, $t(23) = 1.55$, $p > .10$. The repetition main effect resulted because averaged across all item types, the FA rate for blue lures ($M = .29$) was significantly higher than for red lures ($M = .21$). Importantly, results from a paired-samples t test conformed with the observation that the FA rate was greater for blue natural words than for red natural words, $t(23) = 2.73$, $p < .015$, $SE = .032$, $\eta^2 = .245$. Finally, the item x repetition interaction was not significant, $F(2, 46) = 1.27$, $p > .25$.

Table 19. Experiment 8: Mean estimates of *d'* (standard deviations in parentheses) for the three items types, presented in blue and in red ($N = 24$).

	Blue ("studied once")	Red ("studied three times")
Natural	1.50 (.83)	2.28 (.59)
Regular	.62 (.55)	1.56 (.80)
Irregular	.66 (.51)	1.24 (.80)

Discrimination Estimate. Discrimination (*d'*) data (see Table 19) reflected the pattern of findings in hit and FA rates. There was a significant main effect of item, $F(2, 46) = 34.85$, $p < .001$, $MSE = .354$, $\eta^2 = .602$. The superior performance, in terms of hit and FA rates, for natural words over both regular and irregular nonwords was supported by the analysis in *d'*. Post-hoc analyses ($\alpha = .0167$) showed that discrimination was significantly better for natural words ($M = 1.89$) than for both regular nonwords ($M = 1.09$) and irregular nonwords ($M = .95$), $t(23) = 6.20$, $p < .001$, $SE = .130$, $\eta^2 = .626$, and $t(23) = 8.19$, $p < .001$, $SE = .115$, $\eta^2 = .745$, respectively. Discrimination for regular and irregular nonwords did not differ significantly from each other, $t(23) = 1.13$, $p > .25$. Apart from the item main effect, the main effect of repetition was also significant, $F(1, 23) = 41.92$, $p < .001$, $MSE = .507$, $\eta^2 = .646$. Discrimination for items presented in red ($M = 1.70$) was significantly better than

items presented in blue (M = .93). The item x repetition interaction was not significant, $F(2, 46) = 1.04, p > .35$.

6.4.3 Discussion

Overall, in Experiment 8, test items shown in red produced higher hit and lower FA rates than those shown in blue, and this therefore constitutes a replication of the findings from Experiment 7. Further, these results are consistent with a SD model which allows within-list criterion shifts (e.g., J. Brown et al., 1977). First, the increase in hit rates for experimentally strengthened (strong) items could be explained by proposing that the distribution for these targets is located to the right of the distribution for weak targets (see Figure 2, section 5.2). Second, the decrease in FA rates for red items, relative to blue items, can be explained by within-list criterion shifts. By this account, there is one distribution comprising of lures from both the strong and the weak item category. However, because the criterion adopted by participants is more conservative when encountering an item from the strong category than an item from the weak category, the FA rate would be lower for the strong item group than for the weak item group. Psychologically, contrary to Stretch and Wixted's (1998) conclusions, evidence of such within-list criterion shifts suggests that at least in the current paradigm (where item memorability was emphasised by the use of colour cues and labels), participants were able to evaluate an item's expected level of memorability on a trial-by-trial basis, and use this information to make correct rejections.

Procedurally, Experiment 8 differed from Experiment 7 in that the presentation duration for study items was shorter and there was a delay between study and test. A direct consequence of these two procedural changes was that recognition performance of Experiment 8 was impaired in comparison with that of Experiment 7. Descriptively, comparing across the two experiments, the procedural modifications in Experiment 8 were effective in both reducing the hit rate and augmenting the FA rate. This observation was supported by *d'* data, which showed that averaged across item types and repetition/colour, *d'* was significantly lower in Experiment 8 than in Experiment 7, $F(1, 46) = 16.85, p < .001, MSE = 1.086, \eta^2 = .268$. Despite this finding, the general pattern of hit and FA rates produced by red and blue items in Experiment 8 paralleled that in Experiment 7. More importantly, Experiment 8

provided more convincing data suggesting that item memorability (as cued by the colour of the test item) plays a role in the recognition judgment of not only intrinsically unmemorable items like nonwords, but also intrinsically memorable items like natural words. Thus, although natural words (probably due to their meaningfulness) are generally regarded to be intrinsically memorable, extrinsic factors imposed by experimental manipulations could affect the overall memorability level of these items. In turn, it might be that in unfavourable encoding conditions, participants are more inclined to appreciate the utility of using memorability cues (the test item's colour in this case) in reducing the FA rate of intrinsically memorable items.

6.5 Findings from Experiment 7 and 8: Implications on the Hension Effect

By demonstrating within-list strength-based effects on FA rates, Experiments 7 and 8 have bolstered the claim that item memorability is considered by participants in their recognition decisions on a trial-by-trial basis. This argument is especially relevant in explaining how in mirror effects, such as that produced by the hension effect paradigm, the FA rate is lower for one item class (i.e., natural words) than another (i.e., regular nonwords, and also irregular nonwords). From the SD perspective, memorability-based correct rejections have typically been conceptualised in terms of specific criterion placements which are determined by each test item's assessed memorability. Because the presentation order of the items from the three item classes are thoroughly randomised during test, item memorability would therefore vary from trial to trial, and as such, within-list, trial-by-trial criterion shifts are necessary to account for FA rate differences observed across item classes.

6.6 Intrinsic Versus Extrinsic Item Memorability

Overall, the important finding from Experiments 7 and 8 was that FA suppression could result either through high intrinsic memorability, which is dependent on the item's inherent characteristics (e.g., meaningfulness), or high extrinsic memorability, which is dependent on experimental factors (e.g., repetition of study episodes; Benjamin & Bawa, 2004; J. Brown, 1976). In this way, FA rates could be reduced for items deficient in intrinsic memorability if their extrinsic memorability could be enhanced experimentally. However, the findings here hinted

that participants might not readily incorporate information pertaining to extrinsic memorability into their decisional processes, particularly if memorability cues were in some way ambiguous, such as those employed by Stretch and Wixted (1998). It seems that clear explicit cues, implemented on a trial-by-trial basis, are necessary in persuading participants to consider an item's extrinsic memorability during recognition tests. It is perhaps due to this reason that past attempts to generate mirror effects using within-list strength manipulations have failed (e.g., Morrell et al., 2002; Stretch & Wixted, 1998).

The difficulty in producing the mirror pattern, by manipulating item memorability within list, can be contrasted with the prevalence of the mirror effect observed among item groups differing in intrinsic memorability. Apart from the hension effect and the WFE, which relied on linguistic materials (i.e., high and low frequency words, nonwords), paradigms using nonverbal items have also generated the mirror pattern. For example, in the recognition memory of faces, high memorability (as indexed by facial caricaturedness) has been shown to underlie the production of not only hits, but also correct rejections (Deffenbacher, Johanson, Vetter, & O'Toole, 2000; see also Vokey & Read, 1992, 1995). In Deffenbacher et al.'s experiments, item memorability was manipulated within lists in the sense that both caricatured (memorable) and veridical (less memorable) faces were presented at study and at test. A mirror effect was found such that the more memorable caricatured faces produced a higher hit rate and a lower FA rate than did the less memorable veridical faces. These findings, together with the hension effect and the well-established WFE, suggest that when item memorability is determined by intrinsic item-based factors, the mirror effect is a readily observed phenomenon.

The results from Experiments 7 and 8, which demonstrated a mirror effect using a within-list strength manipulation, was therefore notable because it arose through the manipulation of extrinsic memorability, rather than intrinsic memorability. However, it was likely that these results came about because the labelling system used was effective in emphasising item memorability and thereby greatly assisted participants in incorporating such information in their judgments. This also suggests that compared to intrinsic memorability, extrinsic memorability (because it is affected by experimental conditions) might be too difficult for

participants to *monitor* on an item-by-item basis. Supporting this hypothesis is the recent finding from Singer and Wixted (2006) that the salience of the within-list strength manipulation would affect the likelihood of producing the mirror effect. In their experiment, Singer and Wixted presented participants categories of items to study, and used delay to manipulate item strength within list (see also Singer, Gagnon, & Richards, 2002). That is, some of the categories were studied some time before the test phase, whereas other categories were studied immediately prior to the test phase. Presumably, delay between study and test would decrease item memorability, and thus, the hit rate would be lower and the FA rate would be higher for categories studied before the delay than those studied after the delay (and immediately before test). This mirror pattern was indeed found. However, it was only present when the delay was as long as 2 days, and was absent when the delay was only 20 or 40 min. Singer and Wixted's result therefore suggests that participants were able to sufficiently monitor items' extrinsic memorability only when the strength manipulation had been extreme and salient.

In contrast, the prevalence of the mirror effect when intrinsic, rather than extrinsic memorability was manipulated, might indicate that the item's intrinsic memorability might be intuitively apparent to participants. This argument therefore aligns with Ghetti's (2003) suggestion that item memorability might first and foremost be determined by intrinsic, item-based characteristics, which take precedence over extrinsic factors such as study and test conditions.

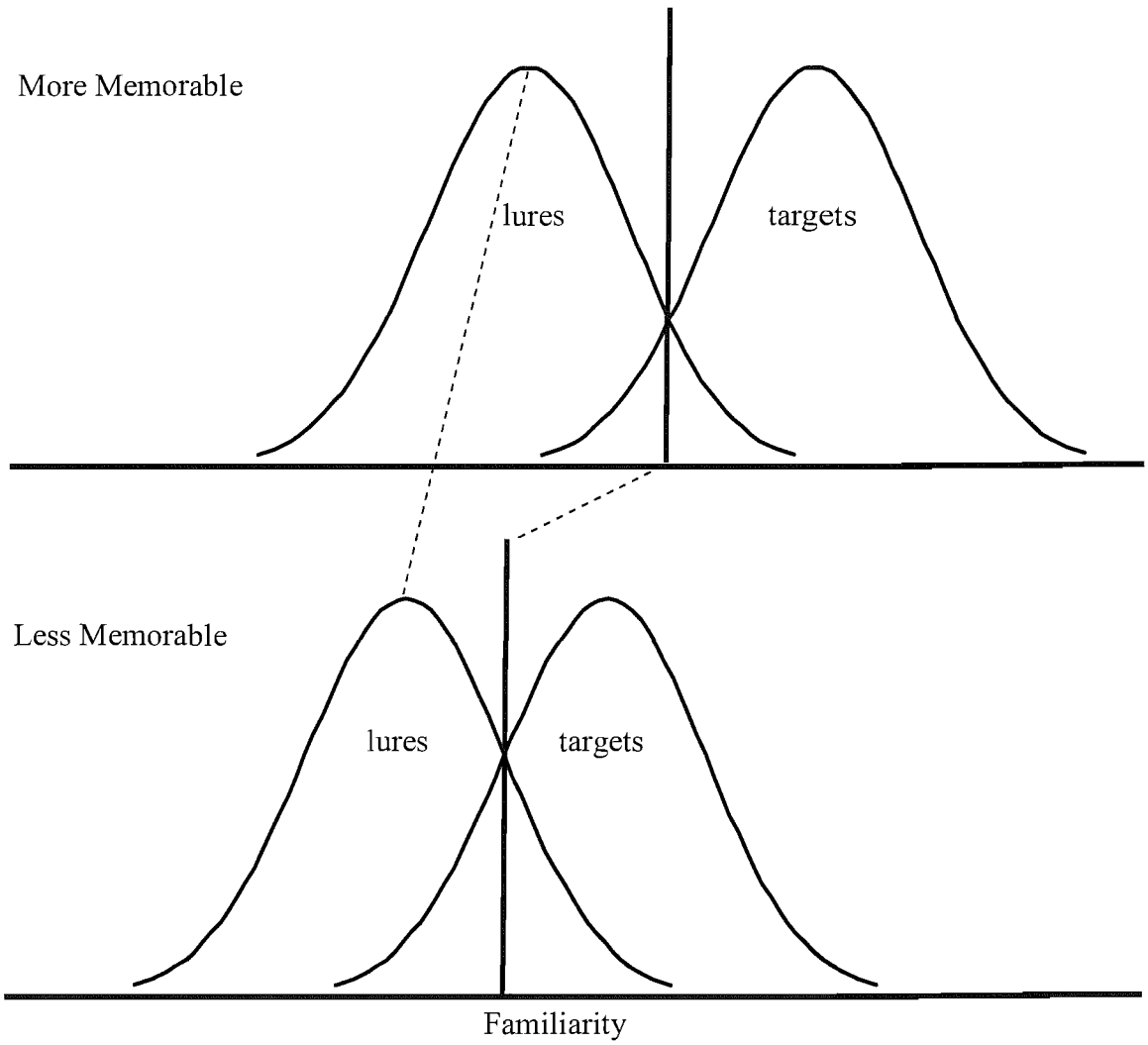
6.7 Modelling the Effects of Memorability: Within-List Criterion Shifts

From the early work by J. Brown et al. (1977) to more recent studies on recognition memory (e.g., Benjamin & Bawa, 2004; Deffenbacher et al., 2004), memorability-based correct rejections have been expressed in terms of criterion shifts in SD models. Thus, the most straightforward interpretation of the within-list strength effects in Experiments 7 and 8, either arising from item-based intrinsic memorability, or from experimentally-manipulated extrinsic memorability, is to assume that the response criterion is adjusted on a trial-by-trial basis. Thus, for a natural word item, or an item presented in red (with the label "studied 3 times or new"), participants would set a conservative response criterion. In contrast, for a nonword item, or an item

presented in blue (with the label "studied once or new"), a more liberal response criterion would be set.

A similar within-list criterion shift model has been proposed recently by Dobbins and Kroll (2005), who obtained a mirror pattern using pictorial materials. These researchers demonstrated that preexperimentally well-known scenes and faces produced a higher hit rate and a lower FA rate than those scenes and faces that were preexperimentally unknown. In Dobbins and Kroll's model, the underlying dimension was assumed to be based on *familiarity*. Familiarity here was used in the sense that participants would have preexperimental experience with familiar (i.e., well-known) rather than unfamiliar (i.e., unknown) items. It followed then that well-known targets *and* well-known lures would form distributions which are located higher on the familiarity dimension than their unknown target and lure counterparts. It should also be noted that in this model, discrimination is better for well-known than unknown items. This is reflected by the way that the distance between the target and lure distributions was larger in the well-known item class than in the unknown item class. As can be seen in Dobbins and Kroll's model (reproduced in Figure 3), the mirror pattern is produced by assuming a more conservative criterion for well-known than for unknown items. In view of these distribution placements on a familiarity-based scale, and that well-known and unknown targets and lures were intermixed in a single test, it was therefore necessary to assume the shifting of the response criterion throughout test.

Figure 3. A within-list criterion-shift model, adapted from "Distinctiveness and the Recognition Mirror Effect: Evidence for an Item-Based Criterion Placement Heuristic," by I. G. Dobbins and N. E. A. Kroll, 2005, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, p. 1186-1198. Preexperimentally known and unknown stimuli were argued to be "more memorable" and "less memorable" stimuli respectively. The setting of the criterion is dependent on whether the stimuli were from a more or less memorable class.



It can be seen that a within-list criterion-shift SD model, such as that proposed by Dobbins and Kroll (2005), is also applicable for the mirror pattern found in the hension effect paradigm. In this case, natural words would be analogous to the well-known items and regular nonwords would be analogous to unknown items. Discrimination data (*d'*) obtained in the previous experiments (e.g., Experiments 6 to 8) indicated that discrimination was significantly better for natural words than regular

nonwords. Thus, like the known items, the separation between targets and lures would be large for natural words; and like the unknown items, this separation would be small for regular nonwords. If the underlying continuum is based on familiarity (in the same sense as defined by Dobbins and Kroll), the placement of the target and lure distributions would also differ across item groups. Natural words, due to their word status, are more familiar than regular nonwords. Hence, the target and lure distributions of natural words would be located further to the right of those of regular nonwords. In this way, by adopting a more liberal response criterion for regular nonwords than for natural words, the mirror pattern is produced for these two item groups[14].

6.8 Arguments Against Within-List Criterion Shift Models

The idea that FA suppression is instigated through criterion shifts, however, has been met with some opposition. One of the objections has centred on the feasibility of trial-by-trial criterion adjustments. As Wixted and Stretch (2000) noted, because items of differing levels of memorability are presented in a fully-randomised order in a standard recognition test, a high number of criterion adjustments is therefore required throughout the duration of the test, which often lasts only a few minutes. It is highly likely that the extra mental energy needed to monitor item memorability and to adjust the criterion accordingly would curtail participants from executing within-list criterion shifts. Even if participants are prepared to adjust their response criterion within a single test, recent work by S. Brown and Steyvers (2005) suggests that such criterion shifts do not occur moment-by-moment, but are slow and require on average, 14 test trials to fully develop.

Debates on the influence of criterion shifts on FA rate levels has also emerged in the research on another false recognition phenomenon, the Deese-Roediger-McDermott (DRM) effect (e.g., Roediger & McDermott, 1995). In this effect, exceptionally high FA rates have been found for "critical lures" – new items that were not presented in the study list, but were strong semantic associates of the studied items. The notion of criterion shifts has been implicated by Miller and Wolford (1999)

[14] Because a mirror pattern is also formed between natural words and irregular nonwords, this within-list criterion shift model can also be applied for these two item classes.

in accounting for this phenomenon. These authors argued that the DRM effect might be caused by participants adopting a liberal criterion when responding to critical lures. However, such a conjecture has since been rejected by a number of researchers (e.g., Gallo, Roediger & McDermott, 2001; see also Roediger & McDermott, 1999; Wickens & Hirshman, 2000, Wixted & Stretch, 2000). For example, Gallo, Roediger and McDermott (2001) argued that because criterion placement is assumed to be under the participant's strategic and conscious control, a reduction in the FA rate of critical lures should be observed if participants were warned before test to adopt a conservative responding strategy. Contrary to this prediction, Gallo et al. (2001) showed that warning about the presence of critical lures did not result in the elimination of the DRM effect.

6.9 Arguments Against Criterion-Based FA Suppression: The Distinctiveness Heuristic

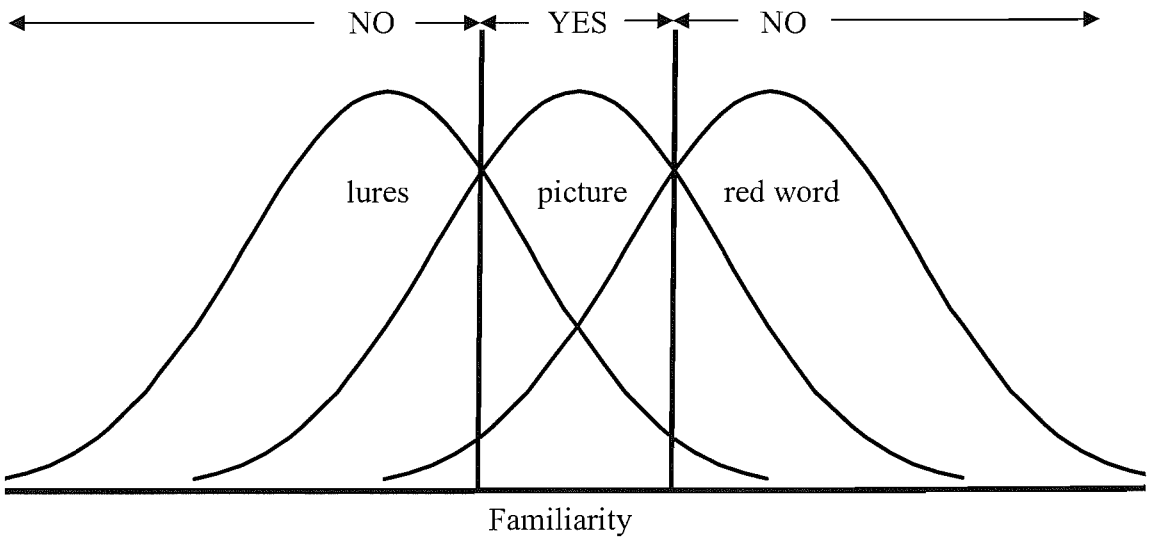
Elsewhere, the involvement of criterion shifts as a FA reduction mechanism has also been questioned in the research on the distinctiveness heuristic. This heuristic has been proposed by Schacter and his colleagues as a metacognitive strategy used by participants to suppress FA rates (e.g., Dodson & Schacter, 2001, 2002; Schacter, Israel, & Racine, 1999). In using this metacognitive strategy, the absence of recollective details for a distinctive (perceptually complex or unique) item is taken as evidence for its prior non-occurrence. An example of how the distinctiveness heuristic is used can be found in Dodson and Schacter (2001). In that investigation, participants either studied the items by saying them aloud, or by hearing them. It was found that the "saying" group produced a significantly lower FA rate than did the "hearing" group. Dodson and Schacter argued that because self-generated (i.e., said aloud) information is generally regarded by participants as more memorable or distinctive than information that was heard (e.g., Johnson, Raye, Foley, & Foley, 1981), retrieval of distinctive memorial information is therefore required before participants in the saying group would judge an item as old. In the absence of such memorial evidence, the item would be readily rejected. Consequently, the use of the distinctiveness heuristic would result in a FA rate suppression in the saying group, relative to the hearing group.

From the SD viewpoint, it has originally been assumed that the use of distinctiveness in rejecting distractors involves adjustments to the response criterion (e.g., Schacter et al., 1999; see also Arndt & Reder, 2003; McCabe, Presmanes, Robertson, & Smith, 2004). That is, criterion setting would be conservative for items belonging to the distinctive condition/category, leading to a suppression of the FA rate. However, this assumption has been more thoroughly examined in a recent investigation by Gallo, Weiss, and Schacter (2004), who devised an experimental paradigm whereby at study, some items were presented as red words and some as pictures. In this paradigm, Gallo et al., rendered red words to be the strong item class, by presenting them repeatedly during study. Picture items were presented only once. After study, participants were given two recognition tests – a *red word test*, and a *picture test*[15]. In the red word test, participants were to give judgments of old only to items studied as red words, and in the picture test, only to items studied as pictures. The effectiveness of the strength (repetition) manipulation on red words was substantiated by the way that red word targets produced a higher hit rate (in the red word test) than did picture targets (in the picture test). Assuming an underlying dimension based on familiarity, Gallo et al. provided a SD-based interpretation of the two tests (see Figure 4). Two response criteria were required in the picture test, such that only items with moderate levels of familiarity (i.e., picture targets) would be accepted. Only one response criterion, however, was needed in the red word test. As could be deduced from Figure 4, more lures should be falsely recognised in the picture test than in the red word test. Contrary to this prediction, Gallo et al. found that the FA rate of lures was still lower in the picture test, suggesting that in this test, participants were strategically seeking distinctive, possibly pictorial information before judging an item as old. These authors concluded that "[e]ven if repetition had made it easier to recall or recollect red words than pictures (a quantitative difference), it is the qualitative difference between the types of expected recollections that is critical for the distinctiveness heuristic" (p.481-482). In relation to the SD model, these researchers therefore argued that the FA suppression seen in the picture test, as a result of the implementation of the distinctiveness heuristic, cannot be easily explained purely by criterion shift mechanisms.

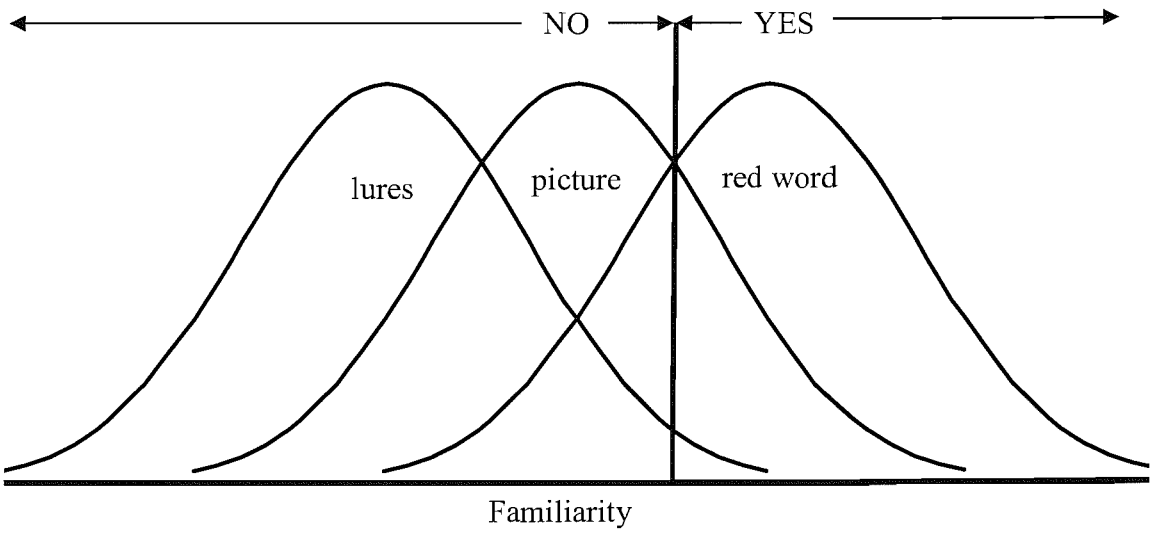
[15] All items at test in Gallo et al.'s (2004) experiment were presented in word form, including items that were studied in pictorial form.

Figure 4. Adapted from "Reducing False Recognition with Criterial Recollection Tests: Distinctiveness Heuristic Versus Criterion Shifts," by D. A. Gallo, J. A. Weiss, and D. L. Schacter, 2004, *Journal of Memory and Language, 51*, p. 473-493. See text for description.

Picture Test



Red Word Test



6.10 Concluding Remarks for Chapter 6

In this chapter, two experiments were reported which provided evidence that item memorability is an important factor in the correct rejection of a lure in recognition tests. In modelling this finding from a SD perspective, it was assumed that a conservative criterion setting for highly memorable items would result in a

suppressed FA rate for these stimuli. Because item memorability is considered for each individual item, the model therefore necessitates the shifting of the response criterion across test trials. Although such a within-list criterion shift model has been proposed to account for similar mirror effects (e.g., Dobbins & Kroll, 2005), other theorists have voiced their scepticism of the plausibility of within-list criterion shifts. Indeed, in the research on the distinctiveness heuristic – another mechanism argued to underlie FA suppression – some investigators have questioned the appropriateness of conceptualising the input of metacognitive processes as criterion shifts in a SD model (e.g., Gallo et al., 2004). In view of these arguments against the notion of within-list criterion shifts, it might be worthwhile to search for an alternative way to represent metacognitive and memorability-based processes in a SD model. Such an alternative, in the form of a multi-process SD model, will be proposed in the next chapter.

Chapter 7

7.1 Distribution Shifts: Separate Distributions for Lures of Differing Memorability

In the final part of the previous chapter, empirical evidence was cited which opposed the use of criterion shifts in modelling metacognitive processes underlying FA rate suppression (e.g., Gallo et al., 2001; Gallo et al., 2004). However, the more compelling theoretical argument against the notion of within-list criterion shifts is that it is difficult to obtain evidence for these occurrences. On the basis of bias estimates (e.g., the parametric measure *C*, cf. Snodgrass & Corwin, 1988), criterion shifts are indistinguishable from distribution shifts (e.g., Roediger & McDermott, 1999; Wickens & Hirshman, 2000, Wixted & Stretch, 2000). As noted by these investigators, bias estimates such as *C* represent the placement of the response criterion relative to the intersection of the target distribution and the lure distribution. As can be seen in Figure 1 (see section 5.1), if the criterion is placed directly at the intersection, responding is said to be unbiased and the estimate would be zero. A negative (liberal) estimate reflects a criterion placement left of the intersection, and a positive (conservative) estimate, right of the intersection. It could be seen that changes in hit and FA rates could occur even if the response criterion is fixed, so long as there are movements in the distributions on the underlying continuum. Moreover, these distribution shifts would manifest rather misleadingly as changes in bias estimates, simply because the relative distance of the fixed response criterion (from the intersection of the target and lure distributions) would also be changed as a result of distribution movements.

Given that distribution shifts, rather than criterion shifts, could bring about hit and FA rates changes, it is perhaps worthwhile to explore whether a SD model based on distribution shifts can account for the mirror effects observed in Experiments 7 and 8. In particular, a multi-process SD model (Wixted & Stretch, 2004) is proposed here where the use of item memorability in FA suppression (i.e., a metacognitive process) is represented by distribution shifts, rather than criterion shifts (see also Tam & Higham, 2006). In this model, lures of differing memorability levels would form separate distributions that are located at distinct points on an underlying, strength-based continuum. To some readers, this idea might seem improbable for the following

reason. In paradigms where extrinsic memorability is manipulated within list, such as when items in each category (e.g., natural words, etc.) were arbitrarily divided into strong (red) and weak (blue) classes (cf. Experiments 7 and 8), red and blue lures should not differ in familiarity or memory strength. This is because lures from neither of the two classes had been presented at study, and as both red and blue lures were essentially from the same item category (e.g., natural words), they should not differ in intrinsic memorability. In this way, lures, regardless of whether they belonged to the strong or weak class, should form a singular distribution.

This assumption of a singular distribution for lures of differing memorability is prevalent in many current SD models, and is probably driven by the belief that because the underlying continuum is *unidimensional*, only *one* process must therefore contribute to its construction. However, as Pastore, Crawley Berens and Skelly (2003) pointed out (see also Wixted & Stretch, 2004), such an assumption does not necessarily hold because according to SD theory, the underlying continuum represents a *decision* dimension, rather than a *process* dimension. That is, it is erroneous to assume that the only one process (e.g., familiarity) can contribute to the value of a test item on this underlying dimension. Instead, by assuming that multiple processes contribute to this decisional value, influences from both metacognitive processes and experimental manipulations on item strength can then be sufficiently modelled.

7.2 Multi-Process SD Model

The multi-process model proposed here is motivated by Wixted and Stretch's (2000, 2004) demonstration that a unidimensional continuum can represent a composite type of strength which is derived from several sources. Wixted and Stretch's multi-process model was put forward in an attempt to reconcile the debate concerning the legitimacy of using SD theory to model remember/know responses in recognition memory (Gardiner, 1988; Tulving, 1985). On the one hand, some theorists in this area have argued that remember and know responses represent qualitatively distinct subjective states, and each is process-pure in the sense that remember responses arise from the recollection of contextual details associated with the study episode, whereas know responses originate from familiarity-based processes (e.g., Gardiner & Gregg, 1997; Gardiner & Java, 1990; Rajaram, 1993; see also Yonelinas, 2002). In contrast, Donaldson (1996, see also Dunn, 2004) suggested that

remember and know responses simply reflect a difference in memory strength, and data from remember/know judgments are compatible with a SD framework. In resolving these two opposing viewpoints, Wixted and Stretch's multi-process model was founded on SD theory, but assumed that memory strength is a composite of strength deriving from both recollection- and familiarity-based processes, rather than from either one of these processes alone. As a consequence, a more appropriate and all-encompassing label – *strength of evidence* – was adopted by Wixted and Stretch for the underlying continuum of their SD model. Thus, the strength of evidence for a studied item (i.e., target) in a recognition test could be represented by the following equation:

$$S_{Target} = S_{Baseline} + S_{Fam} + S_{Ret}$$

where strength of target (S_{Target}) is a summation of baseline level of strength ($S_{Baseline}$, or the item's level of preexperimental strength) and strength from familiarity-based (S_{Fam}) and recollection-based retrieval (S_{Ret}) processes.

Guided by the same principles inherent in Wixted and Stretch's (2004) multi-process model, it is argued here that metacognitively-derived evidence may be represented in a SD model without the implication of criterion shifts (Tam & Higham, 2006). Whereas Wixted and Stretch provided a framework for positive responses (judgments of old in recognition tests), the focus of this multi-process model is on negative responses (judgments of new) and the way memorability-based information is used to promote the correct rejection of lures. In this model, the response criterion is assumed to be fixed, and memorability-based evidence reduces the overall strength of a lure. Specifically, higher item memorability would translate to greater metacognitively-based strength to be subtracted from the lure's overall strength of evidence of prior occurrence. In this way, highly memorable lures would form a separate distribution which is located lower on the underlying continuum than the distribution formed by less memorable lures. Because the criterion is fixed, the FA rate for memorable lures would thereby be lower than that for non-memorable lures. Following Wixted and Stretch (2000, 2004), the strength of evidence of a new item (i.e., lure) could be represented as:

$$S_{Lure} = S_{Baseline} - S_{Metacog}$$

where strength of evidence for a lure (S_{Lure}) is equal to the baseline strength of evidence ($S_{Baseline}$) substracted by $S_{Metacog}$. Here, $S_{Metacog}$ represents strength derived from metacognitive processes (e.g., the distinctiveness heuristic, assessment of extrinsic and intrinsic memorability) and it is substracted from $S_{Baseline}$ because it is taken as evidence *against* the lure's prior occurrence.

It is important to note that unlike previous SD models where the underlying continuum typically represents a strength-based variable, such as familiarity, the continuum for this multi-process model, as in Wixted and Stretch (2004), is labelled as "strength of evidence". In this way, the model conforms with the original principles of SD theory, such that, as argued earlier, this account of recognition memory describes a decision process, not a memory process (Pastore et al., 2003). Hence, the participant's decision during a recognition test is based on the strength of evidence for the item's prior occurrence, and not specifically on a memory strength variable such as familiarity per se.

7.3 Target and Lure Distributions on the Strength-of-Evidence Scale: Evidence for the Multi-Process SD Model

Supporting evidence for the multi-process SD model would be found if groups of new items, which differ in terms of memorability, are shown to form distributions at different points on a "strength-of-evidence" scale. The primary aim of the experiments to be reported in this chapter is to search for such evidence, using materials from the hension effect paradigm. As has been argued in previous chapters, and substantiated by ratings data (Experiment 5, section 4.8), the three item types used in the hension effect paradigm differ from each other in intrinsic memorability. Thus, it was expected that the arrangement of target and lure distributions in this SD model would reflect these memorability assessments. Further, distribution placements were also predicted to be sensitive to manipulations on the item groups' extrinsic memorability. Because it was argued that in the hension effect paradigm, correct rejection for regular nonwords had been inhibited by these items' low intrinsic memorability, the manipulation on extrinsic memorability adopted by Experiments 9 and 10 was directed at this specific group of items. Furthermore, to maximise the enhancement of extrinsic memorability, a different manipulation was used whereby the manipulation was concentrated on only a minority (rather than half, as in

Experiments 7 and 8) of the regular nonword items. Specifically, a minority group of regular nonwords (henceforth denoted as R*+) were presented in purple at study[16]. Other study items – i.e., all natural words (N+), all irregular nonwords (I+), and the majority of the regular nonwords (R+) were presented in black. At test, these studied items were presented in the same colour as in study. For new items, a minority of regular nonword lures (R*-) were tested in purple, whereas the remaining lures – all natural words (N-), all irregular nonwords (I-), and the majority of regular nonwords (R-) were tested in black. In this way, both R*+ targets and R*- lures, unlike their black counterparts, R+ and R-, were enhanced in their extrinsic memorability. Overall, this experimental manipulation has its basis in the *Von Restorff* effect paradigm (see Hunt, 1995), in which memory is enhanced for a "distinctive" study item that is, on a particular dimension, distinguishable from all other study items (which are all similar to each other)[17]. The use of such a paradigm could be more recently seen in the research on memorability-based correct rejections (Ghetti, 2003; Strack & Bless, 1994) and distinctiveness effects in remember/know recognition responses (Kishiyama & Yonelinas, 2003). It is argued here that by presenting a selected minority of regular nonwords in a distinctive colour, their memorability, and therefore overall recognition performance, would be enhanced.

In another departure from the methodology adopted for previous experiments, the effects of memorability on recognition performance would be measured here in a two-alternative-forced-choice (2AFC) test, rather than an old/new recognition test. On each trial in the 2AFC test, two test items will be presented and participants are instructed to choose the item that had been studied. The 2AFC procedure was adopted for two reasons. First, the decision on a 2AFC test trial is presumed to be based on a

---

[16] The plus (+) and minus (-) signs are used here to denote an item's old and new status respectively.

[17] The meaning of "distinctiveness", when used in relation to the Von Restorff paradigm, is therefore different from the sense assumed in the research on the distinctiveness heuristic, where distinctiveness was typically used to refer to the way in which items were encoded (e.g., study items that were said aloud were encoded more distinctively than those that were heard, see Dodson & Schacter, 2001). Although the term has slightly different meanings in these two areas of research, it can be seen that distinctiveness, either generated by encoding conditions (as in the distinctiveness heuristic literature), or by the Von Restorff paradigm, enhances recognition performance by increasing item memorability.

comparison between the two items on a particular dimension. Assuming that this dimension represents the strength of evidence of prior occurrence, the item with greater strength would then be chosen as old. In this way, it is generally regarded that response bias is eliminated in the 2AFC procedure, because the participant need not set a response criterion in order to make the 2AFC decision (Macmillan & Creelman, 2005)[18]. The main rationale for adopting the 2AFC test format, however, is that its criterion-free assumption is consistent with the fixed-criterion feature found in the multi-process model for negative responses, as proposed in the section above. If, contrary to the fixed-criterion assumption, item memorability exerts its effects through criterion shifts, there is no reason to expect that preference (or rejection) rates of lures could be predicted from their memorability levels. As such, on a 2AFC "null" trial that contains a distinctive lure (R*-) and a non-distinctive lure (R-), participants should show no preference for one lure over the other. On the other hand, if, as suggested by the multi-process model, that memorability information is used to determine the strength of evidence of a test item, strength for the distinctive R*- lure would be significantly less than that for the nondistinctive R- lure. Consequently, participants should be more inclined to judge R- as old, and R*- as new.

Similar predictions were made for the comparisons between item groups that were not experimentally manipulated by colour presentation (i.e., those shown in black), but are different in terms of intrinsic memorability. Due to their meaningfulness, natural words are assumed to be intrinsically more memorable than regular and irregular nonwords. In this way, N- (natural word lures) should be lower in strength of evidence than R- (regular nonword lures) and I- (irregular nonword lures).

Although differences among item groups in preference and rejection rates can be determined through analyses of 2AFC data, these differences can be depicted more clearly by ascertaining how target and lure item groups are arranged on a strength-of-evidence scale. In order to construct such a scale, the Thurstonian scaling technique

[18] Technically, some form of bias still exists in the 2AFC procedure, but this bias relates to a tendency to respond according to some other dimension inherent in the experimental context (cf. Hicks & Marsh, 1998). An example of this would be the participant's tendency to choose the item presented on the right rather than the item on the left. Because this type of bias is irrelevant to the point of interest here, bias estimates, as in previous experiments, will not be reported here.

was adopted. An example of how this technique could be used is found in Wixted (1992), who investigated the mirror effect with high frequency, low frequency, and rare words. The attractive feature of this technique is that the constructed scale represents the psychological dimension on which 2AFC judgments are made. In the construction of his scale, which was labelled "subjective sense of prior occurrence", Wixted showed that in accordance to old/new recognition data, rare words did not fit into the mirror pattern established by low frequency and high frequency words (i.e., the WFE, e.g., Glanzer & Bowles, 1976). Rare word targets were placed virtually as high as low frequency targets on the subjective-sense-of-prior-occurrence scale. If rare words conform to the mirror pattern, one would then expect rare word lures to be located as low as low frequency lures at the other end of the scale. Instead, rare word lures were found to be placed higher on the scale than high frequency lures, thus contradicting the mirror effect pattern.

By applying the Thurstonian scaling technique on 2AFC data, Experiments 9 and 10 aimed to establish whether the location of item classes (as targets and lures) on an underlying dimension would follow the predictions made by the multi-process SD model. In these experiments, it was assumed that the unidimensional scale represents a strength-of-evidence variable which receives contribution from both memory-strength-based (e.g., recollection, familiarity) and memorability-based metacognitive processes. Thus, as with the predictions made for the 2AFC data, it was hypothesised that at the higher end (the right-hand side) of the scale, the intrinsically memorable natural word targets (N+) would be located further to the right than the intrinsically unmemorable regular and irregular nonword targets (R+) and (I+). This pattern would be "mirrored" at the lower end (left-hand side) of the scale, with the intrinsically memorable N- lures located further to the left than the intrinsically non-memorable R- and I- lures. Of particular interest in this experiment, however, would be the location of both R*+ and R*- items on this scale. In the Von Restorff effect (e.g., Hunt, 1995), better memory is observed for distinctive items. Hence, it was expected that distinctively-coloured targets (R*+) would be placed higher than non-distinctive R+ items on the strength-of-evidence scale. More importantly, because distinctiveness serves as a form of memorability-based strength that could be subtracted from the overall strength of evidence for lures, it was expected that the distinctively-coloured lures (R*-) would be located further to the left on the scale, compared to the non-

distinctive R- lures. In sum, the outcome from the Thurstonian scaling procedure would elucidate the arrangement of item distributions on a strength-of-evidence scale. Starting from the left-hand side of the scale and moving to the right, the arrangement of item types was predicted to be in the following order: N-, R*-, R-, I-, I+, R+, R*+, N+. In general, supporting evidence for the multi-process model would be found if it could be shown that lures of differing levels of intrinsic and extrinsic memorability form separate distributions on the underlying continuum.

7.4 Experiment 9: Construction of the Strength-of-Evidence Scale (I)

Experiment 9 was conducted in an attempt to construct a strength-of-evidence scale, using the 2AFC and Thurstonian scaling methods as detailed above. In Experiment 9, all study items were presented at the same duration. A minority of regular nonword items were presented in purple such that their distinctiveness would render them to be a highly memorable group.

7.4.1 Method

Participants. Thirty-five students from the University of Southampton participated in return for course credit or £4. For all participants, English was the only language they could speak fluently, and none had participated in other experiments reported in this thesis.

Materials and design. The materials used in Experiment 9 were based on those in Experiments 6 – 8, which were in turn adapted from Whittlesea and Williams (2000). These materials consisted of 40 items from three categories – natural words, regular nonwords, and irregular nonwords (see Appendix F for a listing). However, for the purpose of having enough items to form 2AFC pairs and to create a minority of "distinctively-coloured" regular nonwords (R*), additional items were taken from the original item set provided by Whittlesea and Williams (4 natural words and 4 irregular nonwords), or were specially constructed (18 regular nonwords; see Appendix H for these additional items). These regular nonwords were constructed in accord to the common characteristic of regular nonwords (that they were orthographically regular and could be easily pronounced), while keeping inter-item similarity low. They were similar to the edited set of 40 regular nonwords in terms of length (M = 6.57 characters versus M= 6.38 characters respectively) and bigram

frequency (M = 964.83 versus M = 1015.29 respectively). Thus, in total, there were 146 items in use – 44 natural words, 58 regular nonwords, and 44 irregular nonwords.

Because of the high number of target and lure categories, it was unfeasible to generate a meaningful method of counterbalancing. Random selection was therefore used to determine which items would be studied, and in the case of regular nonwords, which would be presented in purple (i.e., the distinctive colour). For each participant, half of the items from each category were presented during the study phase. Hence, the study phase consisted of 73 items – 22 natural words, 29 regular nonwords, 22 irregular nonwords. Of the 29 regular nonwords, 7 were presented in purple, thus creating a distinctively-coloured minority (i.e., R*+). All other items (all natural words, all irregular nonwords, and the remaining 22 regular nonwords) were presented in black. At test, all 146 items were presented. The 7 regular nonwords that were studied in purple were tested in purple. Seven new regular nonwords were also presented in purple to create a minority of distinctively-coloured lures (i.e., R*-). Because the test phase employed a 2AFC format, 73 test trials can be constructed from the 146 test items (i.e., 146/2 = 73). The majority of these were *standard trials* in that of the pair, one item was a target and one was a lure. Others were *null trials* in that of the pair, both items were targets or both were lures. All the possible types of 2AFC pairings, and the number of trials in each type of pairing are presented in the matrix in Table 20. It can be seen that in pairings where both items were presented in black (i.e., both items were non-distinctive), there were 4 test trials in each of these types of pairings. However, because of the low number of distinctive purple items, there was only 1 trial in each of the possible pairings involving purple items.

Procedure. The procedure for the study phase in Experiment 9 followed closely to that in Experiments 6 – 8 (see section 5.8.5 for details). One exception was that in Experiment 9, study items were presented for 3 s each (participants were informed of this in the study instructions). Participants were also told that while most of the items would be presented in black, a minority of the non-English words would be presented in purple. However, regardless of the colour of the item, they were to try to remember all the items for a later recognition memory test. These instructions were immediately followed by the study phase, and the manner in which this was presented was again identical to that in Experiments 6 – 8. That is, 3 practice items (one from

each category, all shown in black and in a fixed order across participants) were first presented, and these were followed by the 73 study items, presented in a freshly randomised order for each participant. The ITI was 1 s.

Table 20. Construction of 2AFC trials in Experiments 9 and 10. Each cell in this matrix indicates the number of trials in which the row item and the column item were presented together as a pair. N+ = natural word target; R+ = regular nonword target; I+ = irregular nonword target; N- = natural word lure; R- = regular nonword lure; I- = irregular nonword lure; R*+ = distinctive regular nonword target; R*- = distinctive regular nonword lure.

	N+	R+	I+	N-	R-	I-	R*+	R*-
N+	-	4	4	4	4	4	1	1
R+		-	4	4	4	4	1	1
I+			-	4	4	4	1	1
N-				-	4	4	1	1
R-					-	4	1	1
I-						-	1	1
R*+							-	1
R*-								-

At the end of the study phase, participants were given a 10 minute retention interval during which they played a computer game. This retention interval was followed immediately by the test phase. Participants here were informed that on every trial in the test phase, two items would be shown on the screen, one on the left and one on the right. Their task was to decide which item was studied earlier and they were to indicate their choice by pressing either the right key ("p" on the keyboard) or the left key ("q" on the keyboard). As shown in Table 20, there were exactly 4 trials in each matrix cell where both test items contributing to the trial were presented in black (e.g., the N+ /R+ cell). On two of these four trials, items from one item category (e.g., N+) were presented on the left side of the screen and those from the other

category (e.g., R+) were presented on the right. The arrangement was reversed for the other two trials. For the test trials involving one purple item (R*+ or R*-) and one black item, the purple item was presented on the left and right side an equal number of times. In the single trial involving two purple items (the 2AFC trial between R*+ and R*-), the position of the items was randomly determined. In this way, for each participant, targets and lures appeared more or less an equal number of times on both the right and left side of the screen.

Participants were not informed of the presence of null trials. They were, however, told that if studied items were to appear in the test again, the presentation colour would remain the same between study and test. Because of this colouring system, participants were told that there would also be several new non-English items which would be presented in purple, and thus would serve as lures. Following these instructions, the test phase began with 3 practice trials, constructed from the 6 practice items used in previous experiments. All 3 practice trials were standard such that in each, one practice study item (shown at the beginning of the study phase) and one new practice item were presented. All practice items were in black. Immediately after the practice trials, the 73 test trials were presented in a uniquely randomised order for each participant. In each test trial, one item of the 2AFC pair was presented on the left side, and the other item on the right side on the screen. Throughout the test phase, an instruction label was placed at the top of the screen to remind participants that (a) they were to choose the studied item, and (b) the colour of the item was the same between study and test. Apart from these specifications, the procedural details (e.g., font size, ITI, etc.) of the test phase were the same as those in previous experiments (see sections 2.2.1 and 5.8.5 for further details).

7.4.2 Results and Discussion

As detailed earlier (see section 7.3), the 2AFC data gathered in the experiments in this chapter were subjected to the Thurstonian scaling procedure, such that a unidimensional strength-of evidence scale could be constructed. Thus, in this Results and Discussion section, 2AFC data from Experiment 9 will first be reported and discussed. Following this, the implementation of the Thurstonian scaling method will be described, and the resultant strength-of-evidence scale will be presented.

7.4.2.1 Preference Data from 2AFC Trials

The 2AFC data from both standard (in normal typeface) and null trials (in italics) are presented in the matrix in Table 21. Each cell in the matrix represents the proportion of trials, averaged across participants, in which the row item was chosen over the column item as the studied target. It can be seen that pairs of cells that are equidistant from the diagonal are complementary, and therefore the proportions of the two cells necessarily add up to 1.00. For example, the entry in the second column on the first row indicates that averaged across participants, the natural word target (N+) was preferred over the regular nonword target (R+) in .60 of the trials containing these two items. It follows then that the entry in the first column on the second row shows a preference rate of .40 for R+ over N+, and this proportion, together with .60, sum up to 1.00. Similarly, for an example of null trials containing two lures – the preference rate for R*- over N- (.74) and the complementary preference rate for N- over R*- (.26) add up to 1.00. This same principle applies to standard trials containing a target and a lure.

Table 21. Experiment 9: The mean proportions of trials, averaged across participants ($N = 35$), in which the row item was preferred over the column item. Data from standard trials are in normal typeface and data from null trials are in italics.

	N+	R+	I+	N-	R-	I-	R*+	R*-
N+	-	*.60*	*.67*	.92	.81	.76	*.46*	.83
R+	*.40*	-	*.59*	.92	.77	.84	*.49*	.80
I+	*.33*	*.41*	-	.89	.75	.81	*.34*	.89
N-	.08	.08	.11	-	*.19*	*.25*	.06	*.26*
R-	.19	.23	.25	*.81*	-	*.45*	.11	*.57*
I-	.24	.16	.19	*.75*	*.55*	-	.34	*.60*
R*+	*.54*	*.51*	*.66*	.94	.89	.66	-	.86
R*-	.17	.20	.11	*.74*	*.43*	*.40*	.14	-

Two sets of analyses were performed, first on the preference data for targets, and second on the preference data for lures. In each set of analyses, preference rates were first compared among the original item types from the hension effect paradigm (Whittlesea & Williams, 1998) – i.e., N, R, and I items. This was then followed by a second analysis which specifically compared R items with R* items in order to examine the effect of distinctiveness on 2AFC recognition performance.

In the first analysis on target preference rates, the 2AFC data produced by each participant was entered in a matrix identical to that in Table 21. The mean preference rates for N+, R+, and I+ items, for each participant, were then calculated by averaging the proportions across each item row. The means were weighted in that four times more weight was given to cells involving two black items (N, R, I) than to cells involving at least one purple item (R*), as there were four times as many trials in the former than in the latter. The mean preference rates calculated were then analysed in a one-way repeated-measures ANOVA, with item (N+ /R+ /I+) as the within-subjects variable. This analysis showed a significant item main effect, $F(2, 68) = 6.03$, $p < .01$, $MSE = .017$, $\eta^2 = .151$. Post-hoc paired-samples t tests, with Bonferroni adjustments to the alpha ($\alpha = .0167$), revealed that this item main effect was driven primarily by a significantly greater preference rate for N+ ($M = .74$) than for I+ items ($M = .64$), $t(34) = 3.04$, $p < .01$, SE = .035, $\eta^2 = .214$. The difference in preference rates between R+ ($M = .70$) and the I+ items was marginally significant, $t(34) = 2.45$, $p < .02$, $SE = .025$, $\eta^2 = .150$. The comparison between the N+ and the R+ preference rate was not significant, $t(34) = 1.44$, $p > .15$.

To ascertain the effects of distinctiveness on recognition performance, the mean preference rates for R+ and R*+ items were compared. Because colour here was used to manipulate item distinctiveness, a suitable comparison would be between the mean preference rates for R+ and R*+ in trials containing a non-distinctive black item as the alternative. In trials where the R+ item was presented along with another black item, the alternative was one of the five following: N+, I+, N-, R-, and I-. Thus, for each participant, the mean preference rate for R+, averaged over these five trial types was calculated. Similarly, a mean preference rate for R*+ was computed over trials where N+, I+, N-, R- or I- was the alternative. A paired-samples t test was then

conducted on the preference rates for R+ (M = .70) and R*+ (M = .74), which showed no significant difference between the two target types, $t(34) = .886, p > .35$.

As with the analysis on target preference rates, mean preference rates for each lure type from the hension effect paradigm were first calculated for each participant, and these were entered into a one-way repeated-measures ANOVA with item (N- /R- /I-) as the within-subjects factor. A significant item main effect was found, $F(2, 68) = 73.37, p < .001, MSE = .009, \eta^2 = .683$. Subsequent post-hoc paired-samples t tests ($\alpha = .0167$) showed that the mean preference rate for N- (M = .14) was significantly lower than that for R- (M = .38), $t(34) = 11.11, p < .001, SE = .022, \eta^2 = .784$, and for I- ($M$ = .39), $t(34) = 9.19, p < .001, SE = .027, \eta^2 = .713$, with the preference rates for the latter two lure types (R-, I-) not being significantly different from each other, $t(34) = .25, p > .80$.

A paired-samples t test was also performed on the preference rate for R- versus R*-, in order to examine the effects of the lure's distinctiveness on the likelihood that it would be rejected in favour of the alternative item on 2AFC trials. Analogous to the analysis on R+ and R*+ preference rates, the average preference rates for both R- and R*- were derived from trials where the alternative was a black item, i.e., N+, R+, I+, N-, and I-. The t test revealed a marginally significant effect of distinctiveness, $t(34) = 1.79, p < .085, SE = .034, \eta^2 = .086$, suggesting that participants were more inclined to reject a R*- item (M = .33) than a R- item (M = .39).

Overall, the preference rate pattern conforms with previous data from old/new recognition tasks involving items from the hension effect paradigm (e.g., Whittlesea & Williams, 1998, 2000). Generally, natural words produced greater hit rates than did regular nonwords, which in turn produced greater hit rates than did irregular nonwords (see previous experiments reported in this thesis). Consistent with this, the 2AFC data here showed that averaged across trials, N+ was more likely to be chosen as the studied item than was I+, with the R+ preference rate falling in between the two. Because of their meaningfulness, natural words are intrinsically more memorable than the two nonword groups (cf. Experiment 5, section 4.8), as meaning would allow richer contextual details to be encoded. According to Wixted and Stretch's (2004) multi-process model, a target's memory strength is a composite of strength derived

from both recollection- and familiarity-based processes. Presumably then, meaningfulness enhances the overall memory strength of N+ through recollection-based mechanisms. At the same time, N+ also benefits from strength derived from familiarity-based processes because participants have encountered these items preexperimentally. In contrast, R+ items only benefit from familiarity-based strength, as these items possess English-like orthography, for which participants have preexperimental experience. I+ items, however, possess low preexperimental familiarity because their orthography is irregular. In this way, preference rate data are consistent with the hypothesised levels of strength and *intrinsic memorability* for the three item groups.

In contrast, the predicted effects on target preference rates, arising from the Von Restorff-based manipulation on R*+ items, were not found. It was expected that by presenting a minority of regular nonwords in a distinctive colour, the (extrinsic) memorability of these items would be enhanced. Contrary to this expectation, purple R*+ targets did not show a significantly higher preference rate in comparison to black R+ targets (although the difference is numerically in the correct direction). This finding suggests that the Von Restorff paradigm might not be a sufficiently potent manipulation in enhancing extrinsic memorability. This issue will be addressed later in this Discussion section, and in Experiment 10.

The intrinsic memorability of the hension effect item groups was also hypothesised to have an impact on the preference rates for lures. Paralleling the data from old/new recognition tests (cf. previous experiments in this thesis), data from the 2AFC test showed that preference rates were significantly higher for nonword lures (R- and I-) than for word lures (N-). As postulated in this thesis, the low levels of FA rates found for natural word lures in old/new recognition tests might be due to participants' ability to utilise these items' memorability (originating from their meaningfulness) to make a large number of correct rejections. In the 2AFC test format, the same argument applies in that participants were able to use memorability to reject N- in favour of the alternative item on a 2AFC trial. Intriguingly, unlike the preference data for targets, there was a stronger indication of a Von Restorff effect in the lures. The contrast between the preference rates for R*- and R- showed a trend

towards significance, suggesting that a regular nonword lure might be more readily rejected if it was distinctive (R*-) than non-distinctive (R-).

7.4.2.2 Using the Thurstonian Scaling Procedure to Construct a Strength-of-Evidence Scale

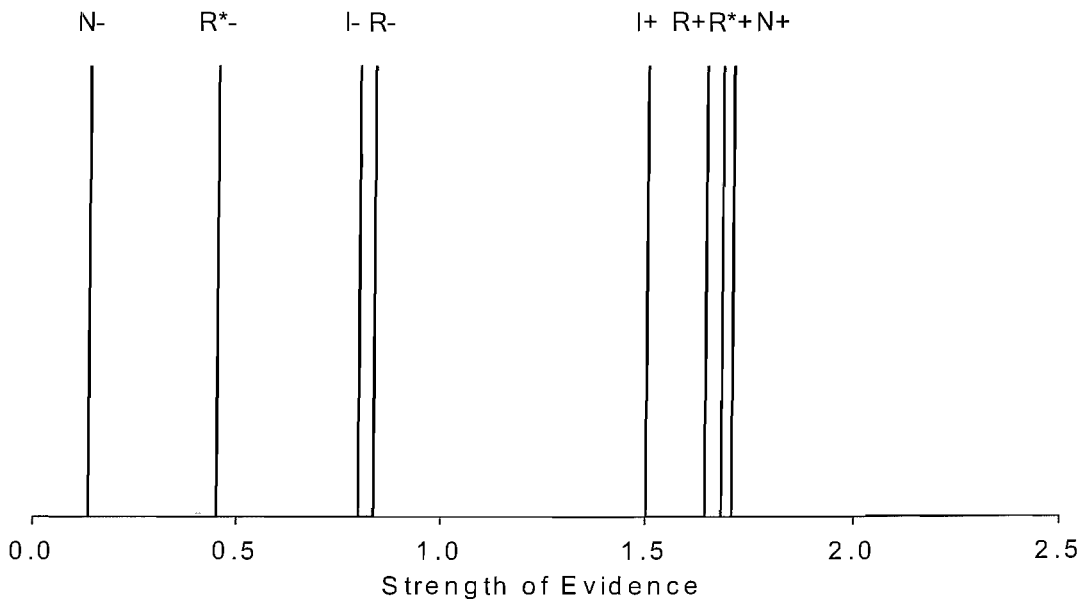
In order to discern how the eight target and lure item groups (N+, R+, I+, R*+, N-, R-, I-, R*-) are arranged on a unidimensional psychological variable (i.e., strength of evidence), the Thurstonian scaling procedure was applied to the 2AFC data obtained in Experiment 9. Because of the assumption that this underlying variable is unidimensional, the eight item types should therefore satisfy the condition of stochastic transitivity. To give a simple example (cf. Coombs, Dawes, & Tversky, 1970), if it is assumed that both the probability of A being preferred over B, and the probability of B preferred over C, are greater than 0.5 (i.e., p(A, B) > 0.5 and p(B, C) > 0.5), then strong stochastic transitivity is held if p(A, C) – the probability of A being preferred over C – is greater than both p(A, B) and p(B, C), that is, p(A, C) > max[p(A, B), p(B, C)]. Moderate transitivity is achieved if p(A, C) is greater than either p(A, B) or p(B, C), that is, p(A, C) > min[p(A, B), p(B, C)].

With eight target and lure groups, there are in total 350 tests for stochastic transitivity, each involving either three or four item groups. For example, in a test involving four groups: p(N+, R+) = .60, p(R+, I-) = .84, and p(I-, N-) = .75, strong stochastic transitivity is held because p(N+, N-) = .92, which is greater than max[p(N+, R+), p(R+, I-), p(I-, N-)]. It was found that 75.14% of the tests satisfied the condition for strong stochastic transitivity, and almost all of the tests (98.57%) showed moderate stochastic transitivity. It was therefore reasonable to assume unidimensionality for the psychological variable on which responding was made.

In implementing the Thurstonian scaling procedure (cf. Baird & Noma, 1978; Wixted, 1992), the 2AFC data (expressed as proportions) were averaged across all participants and were arranged in a 8 x 8 matrix, as shown in Table 21. These proportions were then converted into z-scores, assuming that the variance of both the target distribution and the lure distribution equalled 1. The scale value for each item class (on each row) was then determined by calculating the mean of the z-scores across that row. Again, because of the unequal number of trials contributing to

various cells in the matrix, the means were weighted accordingly. Additionally, following Wixted (1992), an arbitrary value of 1 was added to each scale value to render all scale values positive. The resulting scale values for the eight item classes – N+, R+, I+, N-, R-, I-, R*+ and R*- are: 1.71, 1.65, 1.50, 0.14, 0.84, 0.80, 1.69, 0.45 respectively. The locations of the item classes on the underlying dimension (labelled as "strength of evidence") are graphically presented in Figure 5.

Figure 5. The strength-of-evidence scale constructed for Experiment 9, using the Thurstonian scaling technique: N = natural words, R = (non-distinctive) regular nonwords, I = irregular nonwords, R* = distinctive regular nonwords. The symbols "+" and "–" denote target and lure status respectively.



Unsurprisingly, the four target groups are positioned on the right half, and the four lure groups on the left half of the scale (see Figure 5). This is because evidence of prior occurrence was stronger for studied targets than for non-studied lures. Of greater interest is the specific arrangement among target and lure groups. Overall, the arrangement is consistent with the results from preference rates analyses. The absence of differences in preference rates among target groups is reflected by the way that the four target classes cluster closely on the scale. This close clustering of target groups might also underlie the violations of strong stochastic transitivity observed in the preference data (24.86% of the tests for strong stochastic transitivity failed). In contrast, greater "spread" is observed among the lure groups. Interestingly,

conforming with preference rates analyses, it appears that the effects of memorability are more evident among lure than target groups. The analyses showed that the rejection rate was higher for N- than both R- and I- items, and this is reflected by the notable separation of N- from R- and I- lures on the strength-of-evidence scale. Further, in agreement with the marginally significant difference in the rejection rates between R*- and R-, a clear separation between these two lure types can also be seen on the scale. The evidence therefore suggests that item memorability, manipulated here by distinctiveness, might be used by participants in making 2AFC judgments. Despite this, however, N- is still located lower on the strength-of-evidence scale than is R*-. Similarly, in the null trials consisting of an N- and a R*- item, the N- item was more likely to be rejected (R*- was chosen as old in .74 of the trials, see Table 21). It appears that intrinsic memorability (deriving from the item's meaningfulness) provided a more powerful justification for rejection than did experimentally-enhanced extrinsic memorability (deriving from the distinctiveness of colour). This idea aligns with the argument that intrinsic memorability may take precedence over extrinsic memorability in recognition judgments (Ghetti, 2003), and with the earlier conjecture that extrinsic memorability may exert limited influence on correct rejections because participants may have difficulty in monitoring this factor (see section 6.6).

It is unclear why the effects of item memorability were not produced among the target groups. It may be that three seconds of study duration was sufficiently long enough to permit all targets, including the non-memorable R+ and I+, to be encoded effectively. That is, the benefit of long study duration for all items might have negated the memorability advantage possessed by N+ and R*+ targets. To address this issue, and to replicate the general findings for lures from Experiment 9, Experiment 10 was carried out after slight modifications were made to the experimental paradigm. First, items were studied for a shorter study duration of 1 s each. Past research showed that recognition performance (particularly in terms of hit rates) could be affected by study duration (e.g., Hirshman & Palij, 1992; Malmberg & Nelson, 2003; Ruiz et al., 2004). In relation to the hension effect paradigm, Experiment 6 reported in this thesis (see section 5.8) further demonstrated that effects of study duration on hit rates were influenced by the item's inherent characteristics. Specifically, it was found that the hit rate of irregular nonwords (which are low in intrinsic memorability) decreased significantly from a long (3 s) to a short (0.5 s) study duration (see Table 14, section

5.8.6). This hit rate decrease was not statistically significant in regular nonwords (which have moderate intrinsic memorability), and was virtually absent in natural words (which are high in intrinsic memorability). Put another way, hit rate differences among the three item groups were greater in the short than in the long study duration condition. On this basis, it was predicted that with a shorter study duration implemented in Experiment 10, the mean preference rate for N+ targets would be greater than that for R+ and I+ targets. A clear demarcation was also expected to be found between N+ and the nonword target groups (R+, I+) on the strength-of-evidence scale.

Data from Experiment 9 also showed that the preference rate for R*+ targets was not significantly higher than that for R+ targets. The reason for this null result was unclear since it was hypothesised that R*+, due to the Von Restorff-based manipulation, should be more memorable than R+ targets. To address this issue in Experiment 10, a second procedural modification was implemented whereby R*+ items, apart from being distinctively coloured, were presented three times during study. All other study items were presented only once. Thus, as in Experiment 9, a minority of regular nonwords (R*+ and R*- items) would be presented in purple at study and at test. However, unlike Experiment 9, a strength manipulation, namely repetition, was applied to R*+ targets. In this way, R*+ targets received the same time period in study exposure in both experiments (i.e., 3 s per item), but the exposure in Experiment 10 was divided across three study presentations. It was hypothesised that repetition would sufficiently increase the strength of evidence of R*+, relative to R+ targets, as it has been argued that repetition creates multiple traces to the study episode, thus increasing the likelihood that studied targets are recollected during test (Dewhurst & Anderson, 1999; Hintzman, 1976). In terms of Wixted and Stretch's (2004) multi-process model, repetition would increase recollection-based strength, and therefore the overall strength of evidence for the item. In short, it was predicted that the repeated and hence highly memorable R*+ targets would show a greater preference rate than would the non-repeated (and therefore non-memorable) R+ target, and this pattern of 2AFC performance would be reflected on the strength-of-evidence scale.

The outcome for lure item groups in Experiment 10, on the other hand, was predicted to follow the pattern shown in Experiment 9. N- lures, being intrinsically memorable, would be more likely to be rejected on 2AFC trials than would the intrinsically less memorable R- and I- lures. Because R*+ targets were repeated during study (and were distinctively-coloured), participants would be able to utilise the enhanced memorability of the R* item group to reject R*- lures. Thus, the mean preference rate of R*- was predicted to be lower than that for R- lures. The pattern in these 2AFC data would therefore suggest that on the strength-of-evidence scale, N- and R*- would be located further to the left than would the R- and I- lure groups.

7.5 Experiment 10: Construction of the Strength-of-Evidence Scale (II)

In Experiment 10, an additional strength manipulation (i.e., repetition) was imposed to enhance the extrinsic memorability of the R* item group. The objectives of the experiment, however, remained the same as those in Experiment 9, that is, to examine the impact of intrinsic and extrinsic item memorability on the arrangement of target and lure distributions on a strength-of-evidence scale.

7.5.1 Method

Participants. Forty participants took part in Experiment 10. Of these, 19 were students from the University of Southampton, who were reimbursed £4 each for their participation. The remaining 21 participants were final-year students from a local secondary college, for whom the experiment constituted an activity on the university's "open day". All participants spoke English as their only fluent language, and none had participated in other experiments reported in this thesis.

Materials and Design. These were identical to those in Experiment 9.

Procedure. The procedure of Experiment 10 was the same as in Experiment 9 apart from the following exceptions. First, participants were tested in groups of 1 to 7. Second, all items were presented for 1 s each during study (with an ITI of 1 s). Third, R*+ targets (i.e., regular nonwords which were randomly selected by the computer to be presented in purple) were repeated three times each during the study phase. Participants were informed of this colour and repetition manipulation in the pre-study instructions, and reminded of it later in the pre-test instructions.

7.5.2 Results and Discussion

Preference data from standard and null trials are shown in Table 22. The data are arranged in the same way as described in the Results section of Experiment 9 (see section 7.4.2.1).

As in Experiment 9, a mean preference rate for each target and lure type was established by calculating the weighted average across each row in Table 22. A one-way repeated-measures ANOVA was first conducted on the preference rates of target groups: N+, R+, and I+, which constituted the original item categories in the hension effect paradigm. This analysis (item: N+ /R+ /I+) showed a significant item main effect, $F(2, 78) = 10.10, p < .001, MSE = .020, \eta^2 = .206$. Subsequent post-hoc paired-samples t tests ($\alpha = .0167$) showed that the preference rates for N+ ($M = .72$) was significantly higher than that for R+ ($M = .60$), $t(39) = 3.45, p < .01, SE = .034, \eta^2 = .234$, as well as for I+ ($M = .59$), $t(39) = 3.63, p < .001, SE = .036, \eta^2 = .253$. The comparison between R+ and I+ preference rates was not significant, $t(39) = .506, p > .60$.

Table 22. Experiment 10: The mean proportions of trials, averaged across participants ($N = 40$), in which the row item was preferred over the column item. Data from standard trials are in normal typeface and data from null trials are in italics.

	N+	R+	I+	N-	R-	I-	R*+	R*-
N+	-	*.64*	*.68*	.84	.76	.77	*.20*	.88
R+	*.36*	-	*.54*	.81	.68	.74	*.08*	.70
I+	*.32*	*.46*	-	.81	.75	.67	*.15*	.80
N-	.16	.19	.19	-	*.26*	*.33*	.05	*.55*
R-	.24	.32	.25	*.74*	-	*.47*	.03	*.63*
I-	.23	.26	.33	*.68*	*.53*	-	.15	*.73*
R*+	*.80*	*.93*	*.85*	.95	.98	.85	-	.90
R*-	.13	.30	.20	*.45*	*.38*	*.28*	.10	-

An additional paired-samples t test was conducted on the preference rates of R+ and R*+ to determine the effect of distinctiveness on recognition. As in Experiment 9, these preference rates were averaged across all trials where the alternative item was presented in black (i.e., N+, I+, N-, R-, I-). The t statistic was significant, showing that preference for R*+ (M = .89) exceeded that for R+ (M = .62), $t(39) = 10.66, p < .001, SE = .025, \eta^2 = .744$.

For lure groups (N-, R-, I-), a one-way repeated-measures ANOVA was performed on the preference rate for these items, averaged over all trials. This ANOVA revealed a significant main effect of item, $F(2, 78) = 29.56, p < .001, MSE = .013, \eta^2 = .431$. Follow-up analyses, using paired-samples t tests ($\alpha = .0167$) showed that N- (M = .23) was chosen as old in a significantly lower proportion of trials than both R- (M = .40), $t(39) = 6.33, p < .001, SE = .026, \eta^2 = .507$, and I- lures ($M$ = .41), $t(39) = 6.24, p < .001, SE = .028, \eta^2 = .500$. The comparison between the preference rates of the latter two lures groups (R- and I-) was not significant, $t(39) = .57, p > .55$.

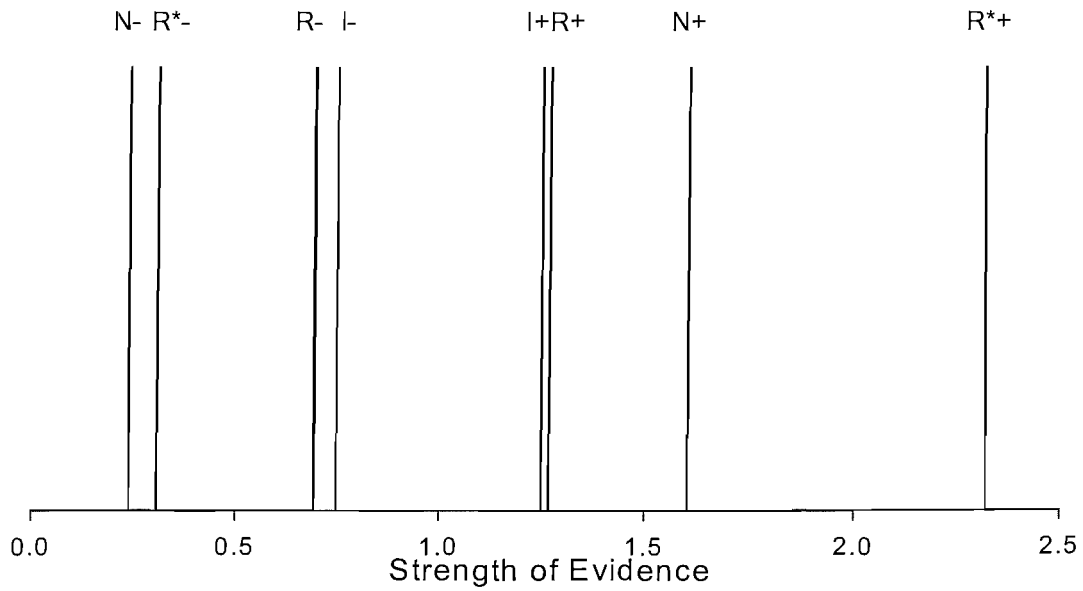
The preference rates for R- and R*-, averaged over trials containing a black alternative item (N+, R+, I+, N- and I-), were compared in a paired-samples t test. This analysis showed that the preference rate for R*- (M = .27) was significantly lower than that for R- (M = .40), $t(39) = 3.49, p < .002, SE = .038, \eta^2 = .238$.

As in Experiment 9, the Thurstonian scaling technique was used on the 2AFC data to construct a strength-of-evidence scale (see section 7.4.2.2 for details of the technique). First, however, tests of stochastic transitivity were conducted to examine the legitimacy of the unidimensionality assumption. The results from these tests were almost identical to those in Experiment 9. Of the 350 tests, 74.86% satisfied the strong stochastic transitivity condition, and nearly all (98.57%) satisfied the condition for moderate stochastic transitivity. Thus, it was reasonable to assume that the eight target and lure item groups could be arranged on a unidimensional scale. This scale, labelled as strength of evidence, is shown in Figure 6. The scale values are 1.61, 1.27, 1.25, 2.32, 0.24, 0.69, 0.75, 0.31 for N+, R+, I+, R*+, N-, R-, I-, R*- respectively.

Most notably, unlike Experiment 9, effects of memorability on target preference rates were found in Experiment 10. Memorable targets (either due to their high level of intrinsic or extrinsic memorability) produced higher preference rates on

2AFC trials (and are located higher on the strength-of-evidence scale) than less memorable targets. In Experiment 9, the study duration of 3 s per item allowed all target types to be encoded sufficiently, thus reducing the advantage in intrinsic memorability that N+ targets have over R+ and I+ targets. With a study duration of 1 s per item in Experiment 10, the preference rate for N+ targets was found to be greater than those for R+ and I+ targets, thus conforming with the hit rate data from Experiment 6 (see section 5.8), which showed greater inter-item differences in hit rates in the short than in the long study duration condition.

Figure 6. The strength-of-evidence scale constructed for Experiment 10, using the Thurstonian scaling technique: N = natural words, R = (non-distinctive) regular nonwords, I = irregular nonwords, R* = (distinctive) regular nonwords. The symbols "+" and "–" denote target and lure status respectively.



Similarly, unlike Experiment 9, R*+ targets produced a significantly higher mean preference rate than did R+ targets. This finding is strikingly represented on the strength-of-evidence scale, where there is a clear demarcation between R*+ targets (located at the extreme right of the scale) and other target types, including the intrinsically memorable N+ targets. This outcome from the scale illustrates the potency of repetition as a strength manipulation on recognition memory. Comparing between the effect of study repetition and of duration on hit rates, Hintzman (1970) found that increases in hit rates were primarily driven by increments in study

frequency rather than in study duration. In the same way, although the overall study exposure of R*+ targets were identical (i.e., 3 s) across Experiments 9 and 10, the repetition manipulation in Experiment 10 appeared to be more effective in increasing the strength of evidence of R*+ targets. It has been argued that repetition allows multiple traces to original study episodes to be formed (Dewhurst & Anderson, 1999; Hintzman, 1976), and thus, from the viewpoint of the multi-process model (Wixted & Stretch, 2004), recollection-based strength (and therefore the overall strength of evidence) of the target item would be increased.

The general findings for lures in Experiment 10 provided a replication of the results from Experiment 9. The effect of intrinsic memorability on the rejection of lures was again evinced by the way N- produced a significantly higher rejection rate (or, lower preference rate) than did the less memorable R- and I- lures. Likewise, the effect of extrinsic memorability on lure rejections was demonstrated in the significantly higher rejection rate for R*- than for R- lures. It is notable that in Experiment 9, when the memorability of R* items was only manipulated by presenting this minority group in a distinctive colour, this same comparison between R*- and R- rejection rates was only marginally significant. Because R*+ items were repeated during study in Experiment 10, presentation colour here could be construed as a memorability cue, used by participants to identify a unique item class (i.e., the strong class) that had been strengthened by repetition. As a result, like the participants in Experiments 7 and 8, participants here were able to use this cue to assess the memorability level expected for lures from the strong class (i.e., R*- lures). In turn, high extrinsic memorability was taken as metacognitively-derived evidence against these lures' prior occurrence.

That repetition rendered particularly compelling evidence to reject R*- could also be seen in the preference data from null trials containing an N- and an R*- lure. In a reversal of Experiment 9's findings, where N- was more likely to be rejected (in .76 of the trials, see Table 21), Experiment 10 showed that R*- was marginally more likely to be rejected (.55 of the trials, see Table 22). Despite this reversal in the preference data, however, the strength-of-evidence scale again showed that N- was located further to the left than was R*-, reflecting the way that across all trials, participants were generally more inclined to reject N- than R*- lures. Thus,

considering the overwhelming preference for R*+ targets in this experiment, it was surprising that this pattern found for targets was not mirrored in the lures data. However, the finding that the strength of evidence for R*- lures was not lower than that for N- lures again suggests the secondary role extrinsic memorability plays (relative to intrinsic memorability) in lure rejections (Ghetti, 2003).

7.6 Concluding Remarks for Chapter 7

In this chapter, two experiments were reported which utilised the 2AFC procedure to examine the effects of item memorability on recognition performance. The 2AFC test format was adopted on the basis of its criterion-free assumption (e.g., Macmillan & Creelman, 2005), an assumption that aligns with the fixed-criterion feature of the multi-process SD model (Wixted & Stretch, 2000, 2004). This model also assumes an underlying dimension representing the strength of evidence of an item's prior occurrence, and allows multiple processes to contribute to this measure of strength. It was proposed here that memorability-based correct rejections may be one of the processes which affects strength, such that high memorability may be construed as evidence *against* a lure's prior occurrence. Consequently, compared to less memorable lures, highly memorable lures would possess lower strength of evidence and be more likely to be rejected. Both Experiments 9 and 10 produced data which support the multi-process SD model, by showing that both intrinsically memorable (N-) and extrinsically memorable lures (R*-) were associated with higher rejection rates on 2AFC trials than were lures of low memorability (i.e., R- and I-).

Using the obtained 2AFC data, the Thurstonian scaling technique was employed to create a clearer graphical representation of how these target and lure types are arranged on a hypothetical strength-of-evidence scale. The scales constructed in both Experiments 9 and 10 demonstrated that different lure types were located at distinct locations on the scale. Highly memorable lure groups (N- and R*-) were placed towards to the lower end of the strength-of-evidence scale, and were separated from less memorable lure groups (R- and I-), which were placed at higher points on the scale. Both the preference data and the strength-of-evidence scale here are therefore consistent with the multi-process SD model, which stipulates that FA suppression for highly memorable lures is a product of distribution shifts, rather than criterion shifts.

Chapter 8

General Discussion

In this final chapter, the findings reported in this thesis will be summarised, and then discussed in relation to the discrepancy-attribution hypothesis, and to the role of memorability in recognition judgments. From a theoretical viewpoint, the latter parts of this chapter will focus on the use of SD-based accounts in modelling data drawn not only from the hension effect paradigm, but other experimental paradigms concerning recognition memory. The chapter will conclude with a discussion of the impact of the current thesis on recognition memory research, and suggestions for future work.

8.1 Manipulations Targeting the Perception of Discrepancy (Experiments 1 – 4)

The ten experiments conducted for this thesis were centred on a specific false recognition phenomenon called the "hension effect" (Whittlesea & Williams, 1998). The effect describes the way that regular nonwords produce a significantly higher FA rates than do natural words and irregular nonwords. To account for this effect, Whittlesea and his colleagues (e.g., Whittlesea, 1997; Whittlesea & Williams, 1998, 2000, 2001a, 2001b) proposed the discrepancy-attribution hypothesis, which stipulated that elevated levels of old judgments (reflected in both hit and FA rates) for regular nonwords arose from the perception of discrepancy experienced during the processing of these stimuli. This sense of discrepancy would in turn trigger an attributional process whereby the fluency associated with stimulus processing would be attributed to a source most plausible to the participant, which, in the context of a recognition test, could be due to the possibility that the item had been studied.

The hension effect had been assumed to be a cornerstone of the discrepancy-attribution hypothesis, and one of the objectives of this thesis has been to examine the validity of this assumption. To achieve this objective, the first four experiments in this thesis were conducted where manipulations were devised to target the sense of discrepancy experienced for a stimulus in the hension effect paradigm. In Experiments 1 and 2, feedback on the test item's processing fluency was made available to participants, in an attempt to modify their assessment of the fluency

perceived for the item. Of specific interest was whether feedback suggesting nonfluent processing for regular nonwords, and fluent processing for irregular nonwords, would respectively eliminate and generate feelings of discrepancy for the two item groups. When feedback was given in the form of exact duration of item pronunciation, no effect of feedback was found (Experiment 1). However, some indication that feedback had an impact on fluency assessment, and hence recognition judgments, was found in Experiment 2, where the feedback was given in a more concrete form of speed description labels (i.e., "fast", "average", "slow"). As was discussed in Chapter 2 (see section 2.3.3), feedback might not necessarily have affected the perception of discrepancy. Indeed, because feedback of "fast" was associated with higher hit and FA rates than feedback of "slow" across all three item types, the effect of feedback appeared to be universal and was not influenced by perceptions of discrepancy.

In Experiment 3, processing fluency was more directly targeted by imposing a different processing task prior to the recognition judgment on each test trial. Specifically, it was reasoned that the resemblance to real English words would render responding to regular nonwords to be particularly nonfluent in a lexical decision task (LDT). In contrast, this lexical judgment would be comparatively easy for irregular nonwords, as they do not resemble any real English words. Consistent with this speculation, and reversing the trend shown in the pronunciation duration data from Experiments 1 and 2, response latency in the LDT was found to be more fluent for irregular nonwords than for regular nonwords. On the basis of the LDT data, it was expected that discrepancy would be eradicated for regular nonwords, but created for irregular nonwords. It followed then that a reversal of the hension effect was predicted. Contrary to this, however, the hension effect was produced in the FA rates, even in an experimental group where the presentation duration of the test item was shortened in order to encourage participants to carry out the LDT on the basis of the item's orthography, rather than pronunciation. Moreover, the hension effect was generated even when participants were not required to perform any preceding task prior to the recognition judgment.

In Experiment 4, further doubts were cast on the relevance of the discrepancy-attribution hypothesis in relation to the hension effect. In this experiment, discrepancy

was manipulated through the creation of meaning for regular and irregular nonwords. According to the discrepancy-attribution hypothesis, meaningfulness would be consistent with regular nonwords' fluent processing, but discrepant with the irregular nonwords' nonfluent processing. That is, if items were given meaning at test, discrepancy should not be experienced for regular nonwords, whereas it should be experienced for irregular nonwords. Consequently, the hension effect (or more specifically, the FA rate difference between regular and irregular nonwords) should be reduced, or even reversed. As with Experiments 1 – 3, such evidence, which would support the discrepancy-attribution hypothesis, was not found.

8.2 The Mirror Effect and Item Memorability (Experiment 5)

Overall, findings from Experiment 1 – 4 failed to produce convincing evidence that the discrepancy-attribution hypothesis is a valid theory for the recognition performance observed in the hension effect paradigm. In searching for an alternative account for the effect, it was observed that the recognition performance between regular and irregular nonwords formed a concordant pattern, with both the hit and FA rates being higher for regular than irregular nonwords. In contrast, the recognition performance produced by natural words and regular nonwords conformed with the "mirror effect", such that the hit rate was higher and the FA rate was lower for natural words than for regular nonwords. Based on these observations, it was postulated that recognition judgments for nonwords (both regular and irregular) might primarily be driven by fluency-based familiarity processes, whereas recollection-based processes might play a more important role in the recognition judgments for natural words. In support of this hypothesis were findings showing that the use of fluency as a basis for recognition judgments may be under participants' strategic control, and may therefore be dependent on experimental conditions (e.g., Westerman et al., 2002) and test items' characteristics (such as their lexicality, e.g., Johnston et al., 1985; Johnston et al., 1991).

Another relevant issue that needed to be addressed, however, was the low FA rate achieved for natural words, relative to regular nonwords. More specifically, although the high hit rate observed for natural words had been attributed to the reliance of recollection-based recognition for these items, it was less clear as to how a large proportion of lures from this item group could be easily rejected. It was argued

that the meaningfulness of natural words rendered these items to be highly memorable, and on the basis of this high memorability, compelling memorial evidence for the item's prior occurrence would be demanded by participants, and when such evidence is absent, the item can be confidently rejected (e.g., J. Brown et al., 1977; Ghetti, 2003; Strack & Bless, 1994). In support of this conjecture, Experiment 5 was conducted which demonstrated that natural words were indeed assessed to be more memorable than both regular and irregular nonwords. Furthermore, the findings from Experiment 5 also suggested that memorability ratings might align more closely to recognition performance (indexed by the discrimination estimate, *d'*) when these ratings were collected during the test phase, rather than in the pre-test study phase or in a post-test context (Benjamin, 2003).

Due to the randomised nature of item presentation in a standard recognition test, the notion of memorability-based correct rejections hinges on an important assumption – that item memorability is assessed individually for each test stimulus. In this way, when a test item is assessed to be highly memorable, convincing memorial evidence is needed before an "old" response is given. In signal-detection (SD) terms, the use of this metacognitive strategy is modelled by the criterion-shift account (J. Brown et al., 1977; see also Stretch & Wixted, 1998), where a conservative response criterion is assumed to be adopted by participants in responding to highly memorable items. Because the metacognitive strategy is selectively applied to only highly memorable items within the test phase, continual, within-list criterion adjustments are also assumed (hence the name "criterion-shift" model). The claim that memorability underlies the FA suppression observed for natural words would therefore be boosted by evidence showing that such within-list criterion shifts can occur. Experiments 6 – 8 were specifically designed for the purpose of obtaining such evidence.

8.3 Memorability-Based Rejections of Lures in a Within-List Context (Experiments 6 – 8)

The paradigms used in Experiments 6 – 8 followed closely to those devised by Wixted and his colleagues (Morrell et al., 2002; Stretch & Wixted, 1998), where item memorability was manipulated within-list. In these experiments, whether the lures belonged to a memorable (strong) or unmemorable (weak) class was distinguished by cues or category membership at test. In Experiment 6, presentation duration (3 s

versus 500 ms) was used to manipulate item memorability between natural words and nonwords (regular and irregular nonword groups combined). It was assumed that at test, participants would be able to identify, using lexicality as a cue, those items belonging to the strong group, and thereby adopt a conservative criterion in order to suppress FA rates for these items. However, for each item type, FA rates did not differ between the strong (long study duration) and the weak (short study duration) conditions, even when, as in the case of irregular nonwords, there were clear strength- or duration-based effects on hit rates. As in Morrell et al., Experiment 6 did not provide evidence supporting within-list criterion shifts, and therefore memorability-based correct rejections.

It was noted, however, that because of the wordlikeness of regular nonwords, participants might not have been able to utilise item lexicality as a memorability cue effectively. A similar criticism on the effectiveness of memorability cues was put forward in relation to Stretch and Wixted's (1998) paradigm, where item memorability was indicated by the stimulus's presentation colour at test. It was argued that even when lures were designated by colour as belonging to the memorable (strong) item class, it was *not explicitly specified* that these items were either presented multiple times or not at all during study. Consequently, participants might entertain the possibility that these items might have been studied once, rather than multiple times, and therefore should be judged as old. The ambiguity surrounding the cueing system, and perhaps subsequent distrust from participants in regards to the reliability of memorability cues, might have precluded memorability-based effects on FA rates from emerging. In view of this, a more explicit labelling system was implemented in Experiments 7 and 8, whose experimental designs were essentially identical to that in Stretch and Wixted (1998, Experiment 5). However, unlike Stretch and Wixted, each test item was accompanied by a decision label which eliminated the potential confusion concerning the memorability cueing system. With the labels in place, the data indicated that participants were able to utilise memorability-based information to reduce the FA rates of items whose memorability had been experimentally enhanced. Furthermore, findings from Experiment 8 demonstrated that given more taxing experimental conditions, the use of item memorability in FA suppression was observed even for items that were, due to their inherent characteristics, intrinsically memorable (e.g., natural words).

It may be useful, therefore, to make a distinction between intrinsic memorability, which is based on preexperimentally-determined item characteristics, and extrinsic memorability, which is affected by experimentally-based factors. It was also noted that in general, previous research has failed to obtaining the mirror effect through manipulations of extrinsic memorability (e.g., Morrel et al., 2002; Stretch & Wixted, 1998). This might in turn suggest that an item's extrinsic memorability might be particularly difficult for participants to monitor. Further, given the prevalence of the mirror effect observed among items of differing intrinsic memorability, it may be that intrinsic, rather than extrinsic, factors is the chief determiner of whether memorability-based correct rejections would occur.

8.4 Inter-Stimulus Similarity and the Hension Effect: Implications for the Discrepancy-Attribution Hypothesis

Experiments 6 – 8 were also characterised by a modification to the makeup of the three item categories. Cleary et al. (2005) recently demonstrated that the high FA rates of regular nonwords might be partly due to the way that many items from this category resemble other items in the stimulus pool. On this reasoning, one third of the items from each item category were removed in order to first, reduce inter-stimulus similarity, and second, equalise item lengths across item categories. Replicating findings reported by Cleary et al., this amendment to the materials resulted in the elimination of the FA rate difference between regular and irregular nonwords. It is difficult to see how the revision to the stimulus pool could affect the sense of discrepancy which is, as hypothesised by Whittlesea and Williams (1998, 2000), associated with the processing of regular nonwords. The partial dissipation of the hension effect, seemingly arising from a decrease in inter-stimulus similarity, was therefore particularly damaging to the claims that the high FA rate of regular nonwords is induced by discrepant, or surprising fluency. Further, along with the failure to eliminate the hension effect through manipulations to discrepancy (Experiments 1 – 4), it appears that the discrepancy-attribution hypothesis might not be suitable in explaining the hension effect.

8.5 Multi-Process SD Model (Experiments 9 – 10)

Findings from Experiments 7 and 8 were consistent with the criterion-shift SD model which assumes a trial-by-trial adjustment of the response criterion. That is, a conservative criterion is set for memorable items encountered during test, whereas for less memorable items, the criterion setting is more liberal. It was noted that similar criterion-based accounts had been proposed for other metacognitive strategies used to suppress false alarms. One such example is the distinctiveness heuristic, a strategy argued to be utilised when failure to retrieve recollective details for a distinctively-encoded item is taken as evidence for its prior non-occurrence (e.g., Dodson & Schacter, 2001, 2002; Schacter et al., 1999; Schacter, Cendan, Dodson, & Clifford, 2001). Originally, it had been presumed that the distinctiveness heuristic operates via criterion shifts – that is, when items had been encoded distinctively, participants would set a conservative criterion during test, resulting in a suppression of the FA rate (see also Arndt & Reder, 2003; McCabe, Presmanes, Robertson, & Smith, 2004). However, a recent investigation by Gallo et al. (2004) has cast doubt on the involvement of criterion shifts underlying the use of the distinctiveness heuristic. Generalising from Gallo et al's conclusions, and considering the dubiousness of obtaining evidence for criterion shifts through bias estimates (e.g., *C*, see section 7.1), it might be necessary to seek out an alternative SD model where metacognitive processes are not equated with criterion shifts. The foundation of this alternative model lies in the multi-process account proposed by Wixted and Stretch (2000, 2004). In this account, an item's value on the underlying axis in the SD model is a composite of strength from various sources, rather than from one single process such as familiarity or perceptual fluency (Pastore et al., 2003). According to this model, memorability-based evidence could be construed as metacognitively-derived strength that is subtracted from the overall "strength of evidence" for the item's prior occurrence. It follows then that lures of high assessed memorability would form a distribution which is lower on the underlying continuum than would lures of low assessed memorability. Assuming a fixed response criterion, the FA rate would necessarily be lower for memorable than for unmemorable lures. In this way, the contribution of metacognitive processes in lure rejection is expressed in terms of *distribution shifts*, rather than criterion shifts.

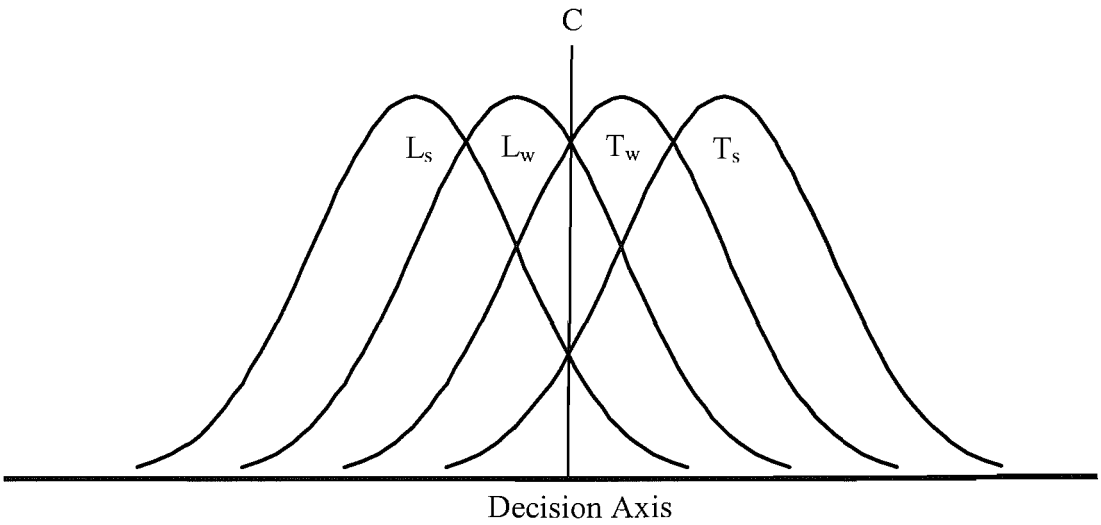
In an attempt to find evidence for this multi-process SD model, a 2AFC test format was adopted in Experiments 9 and 10, with the specific purpose to ascertain the distribution arrangement for targets and lures (from the item types in the hension effect) on a strength-of-evidence scale. The preference data from these experiments showed a significantly higher rejection rate for lures belonging to a class high in either intrinsic memorability (e.g., natural words) or extrinsic memorability (e.g., regular nonwords that had been repeated during study). Using the Thurstonian scaling method on the 2AFC data, a strength-of-evidence scale was constructed which illustrated the hypothetical distribution placements of the target and lure groups from the hension effect paradigm. Consistent with the preference data, lures of differing intrinsic and extrinsic memorability were found to occupy at distinct points on the scale. This finding has therefore greatly bolstered the proposal of using a multi-process SD account to model memorability-based strategies in lure rejections.

8.6 Likelihood-Ratio Models

Some readers may observe a resemblance between the multi-process SD model proposed here and the attention-likelihood theory (ALT; e.g., Glanzer & Adams, 1985, 1990; Glanzer et al., 1993; Glanzer & Bowles, 1976). The model based on ALT – probably the first example of a "likelihood-ratio" model – was directly tailored for the mirror pattern, such as that formed by high and low frequency words in the WFE. It is characterised by a fixed response criterion and specific arrangements of target and lure distributions on the underlying continuum (see Figure 7). Like the multi-process account, Glanzer's likelihood ratio model assumes separate distributions for targets, as well as lures, that are from different item classes. In another similarity to the multi-process account, Glanzer's model maintains that the underlying continuum is a decision axis. It is this crucial feature which allows the separation of lure distributions – a feature that is absent in criterion-shift models (see Figure 2, section 5.2). According to ALT, the item's statistical probability of being old (i.e., likelihood ratio) is assessed, rather than its "memory strength" per se. Applied to the WFE, Glanzer and his colleagues argued that for lures, low frequency words are judged to have a lower likelihood of being old than high frequency words, whereas the reverse is true for targets, thus producing a mirror effect. In this model, memorability-based information is assumed to affect the assessed probability of the

item being new, which in turn impacts on the likelihood ratio that is calculated for the item.

Figure 7. The likelihood-ratio model for the mirror effect, as proposed by Glanzer and his colleagues (e.g., Glanzer & Adams, 1985, 1990; Glanzer et al., 1993). There are separate distributions for items belonging to the strong and weak classes, for targets (T_s and T_w respectively) and for lures (L_s and L_w respectively). The response criterion (C) is fixed.



The model proposed by Glanzer and his colleagues is marked by a unique regularity – that movements of distributions are assumed to be symmetrical. That is, within an item class, if the target distribution is shifted in one direction, the corresponding lure distribution shifts in the opposite direction (e.g., Glanzer et al., 1993). The term "concentering" refers to cases where distributions move towards the midway point on the decision axis, creating greater overlap and less distance between target and lure distributions. In other words, concentering occurs when recognition performance is impaired (as also reflected by a decrease in *d'*). In contrast, if recognition performance is enhanced, the target and lure distributions move away from the midway point on the decision axis, creating less overlap and greater distance between the distributions (and hence an increase in *d'*). This pattern of movement is known as "dispersion" (Glanzer et al., 1993; Hilford et al., 1997).

Through the principles of concentering and dispersion, a SD model based on ALT *necessarily* predicts the full mirror effect (e.g., Glanzer et al., 1993). It is this inflexibility of the original likelihood-ratio model which renders the model difficult to

be reconciled with data showing partial mirror effects (e.g., Morrell et al., 2002; Stretch & Wixted, 1998; see also Hirshman & Arndt, 1997; Hirshman & Palij, 1992; Shiffrin et al., 1995). It follows then that this type of likelihood-ratio model, as envisaged by Glanzer and his colleagues, cannot provide an adequate account for data from a large number of studies, including Experiment 6 in the current thesis, where within-list strength manipulation resulted only in hit rate changes, but not FA rate changes. In the same way, the ALT cannot accommodate findings showing the elimination of the hit rate component of the mirror effect, while the FA rate portion of the effect remained intact (e.g., Hirshman & Arndt, 1997).

However, more recent likelihood-ratio models such as the REM (Retrieving-Effectively-from-Memory) model (Shiffrin & Steyvers, 1997; see also the Subjective-Likelihood Theory proposed by McClelland & Chappell, 1998)[19], have incorporated elements which remedied the inflexibility of Glanzer's ALT account. As such, the newer likelihood-ratio models can therefore accommodate the partial mirror effects obtained in past research (e.g., Stretch & Wixted, Experiment 6 in this thesis). For instance, in the REM model, each item in a recognition test is said to consist of *an array of features*, with each feature being represented by a numerical value. For example, a study item might be represented by this array: {1, 2, 3, 3, 2, 1}. A representation of the study item is stored during each study episode, and this representation is called an *image*, which is also expressed as an array of features. The image is assumed to be imperfect, but with study repetitions, the values of the features will become more aligned with those of the study item. For example, the image corresponding to the above study item {1, 2, 3, 3, 2, 1} might be {1, 0, 0, 3, 0, 1} after the first study episode. After several repetitions of the study item, however, the image is expected to become more accurate: {1, 2, 0, 3, 2, 1}.

On a test trial, the test item's array of feature values will be compared with those contained in *each individual* image stored. For each comparison, the model calculates: first, the likelihood that the features match (or mismatch), given that the test item was a target, and second, the likelihood that the features match (or

[19] As acknowledged by Shiffrin and Steyvers (1997), their likelihood-ratio model, and that put forward by McClelland and Chappell (1998) share a large number of similarities, even though both models were developed independently. Because of this, only the REM model will be described (briefly) here.

mismatch), given that the test item was a lure. The ratio of the two computed likelihoods is the likelihood ratio. For the REM model, Shiffrin and Steyvers (1997) have arbitrarily set the model to respond "old" if this likelihood ratio exceeds 1.0.

The REM model predicts full mirror effects for items of differing intrinsic characteristics (such as in the WFE), but does not always predict the mirror pattern when the strength manipulation is imposed within list. For example, Morrell et al. (2002) performed a REM simulation which produced results conforming with the partial mirror effect they obtained in their within-list strength-manipulation experiments (where one semantic category, e.g., profession words, was selectively repeated, relative to words from another semantic category, e.g., location words; see section 5.8.2). However, this partial mirror pattern was only produced by the REM model if *non-preferential global matching* was assumed. That is, the partial mirror pattern was achieved only if every image, regardless of its category membership (i.e., every profession and location word stored in memory), was compared with the given test item. In contrast, if the matching process was assumed to be *preferential* (i.e., the test item was compared with only stored images in the same semantic category), the full mirror pattern would result (Morrell et al., 2002).

The outcome from Morrell et al.'s (2002) REM simulation suggests that in their experiments, where strength was manipulated within list, the matching process performed by their participants was non-preferential rather than preferential, as a partial mirror effect, rather than a full mirror effect, was produced by their human participants. Similarly, the findings from Experiments 7 and 8 (see Chapter 6), suggest that preferential matching process was performed by participants here, as the mirror pattern was observed between strong (red) and weak (blue) items. It might be that in the context of a salient labelling system, participants were able to carry out preferential, rather than nonpreferential matching. Thus, the challenge for current likelihood-ratio models is to incorporate a term which embodies the influence of extrinsic memorability cues. In their present state, likelihood-ratio models express items as arrays of features based only on the lexical/semantic characteristics of the item. A term is therefore needed in the model which will specify that for a test item assessed (or indicated by extrinsic cues) as high in memorability, a high number of features are expected to be matched between the item and a stored image. Insufficient

matches would consequently hold great diagnostic significance to reflect that the test item was a lure, rather than a target.

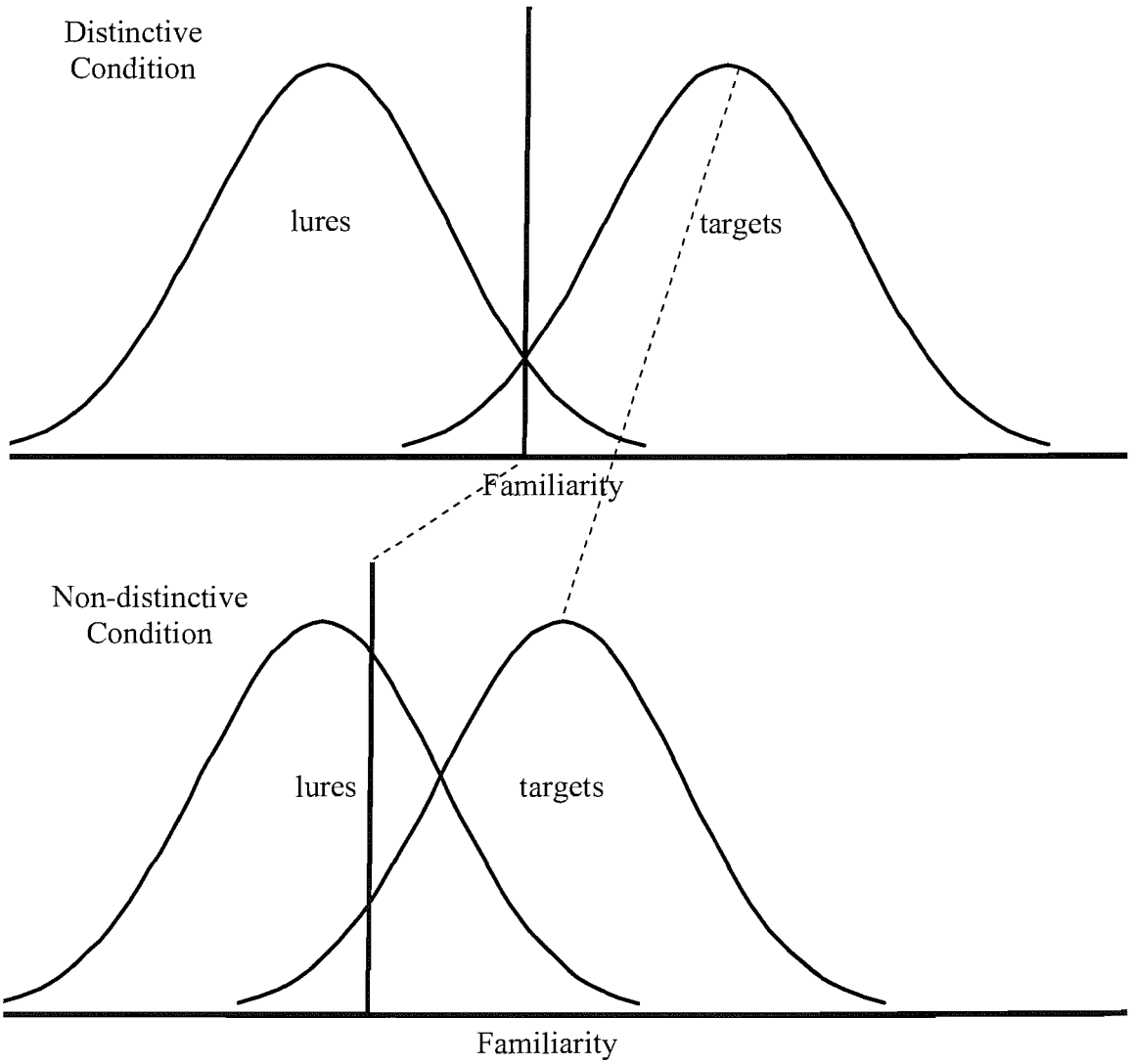
8.7 Comparisons Among Criterion-Shift, Likelihood-Ratio, and Multi-Process Accounts: Modelling the Distinctiveness Heuristic

There are clear parallels between the conjecture of memorability-based correct rejections, as proposed originally by J. Brown et al. (1977), and the use of the distinctiveness heuristic as a FA suppression mechanism, as put forward more recently by Schacter and his colleagues (e.g., Schacter et al., 1999). The former hypothesis assumes that for a test stimulus deemed to be highly memorable, strong memorial evidence for its prior occurrence is required before an "old" response is warranted. In the latter hypothesis, the absence of recollective details available for a highly distinctive test stimulus was conducive to a "new" response. These two proposed metacognitive mechanisms may be "two sides of the same coin", as both are strategies deployed by participants to reduce FA rates by increasing the number of correct rejections made. Given the close relationship between these two metacognitive processes associated with the recognition judgments of lures, any potential SD model for memorability-based correct rejections should therefore also be capable of accommodating the findings from the distinctiveness heuristic literature.

As detailed earlier (see section 6.9), recent empirical evidence from Gallo et al. (2004) indicated that the distinctive heuristic is unlikely to operate via criterion shifts. At the same time, a peculiarity in the research on the distinctiveness heuristic may also pose problems for the criterion-shift model. It has typically been found that the use of distinctiveness heuristic affects FA rates only, with minimal impact on the hit rates. For example, in the investigation conducted by Dodson and Schacter (2001; as described earlier in section 6.9), suppression of FA rates was found in the participant group who encoded the study items by saying them aloud, relative to the group who only heard the items during study. However, in terms of hit rates, there was no significant difference between the two groups. The same pattern of results (stable hit rates and FA rate differences) was found by Schacter and his colleagues using word versus picture encoding conditions (Dodson & Schacter, 2002; Schacter et al., 1999; Schacter et al., 2001), and by other researchers using encoding conditions of differing distinctiveness (e.g., Arndt & Reder, 2003; Budson, Dodson, Daffner, &

Schacter, 2005; Budson, Dodson, et al., 2005; Budson, Droller, et al., 2005; Ghetti, 2003; Ghetti & Qin, & Goodman, 2002; Kishiyama & Yonelinas, 2003; Kishiyama, Yonelinas, & Lazzara, 2004; McCabe et al., 2004; Smith & Hunt, 1998; Strack & Bless, 1994).

Figure 8. A hypothetical criterion-shift model for the distinctiveness heuristic. Note that the area under the target distribution, and to the right of the criterion, is identical in both conditions to indicate equal hit rates. The movement of the criterion therefore has to be exact in order to maintain this stable hit rate pattern.



If the use of the distinctiveness heuristic were to be conceptualised in terms of criterion shifts (e.g., Brown et al., 1977), a hypothetical model is shown in Figure 8. In this model, targets encoded in the distinctive condition would be higher in memory strength (e.g., familiarity) than those encoded in the non-distinctive condition, and

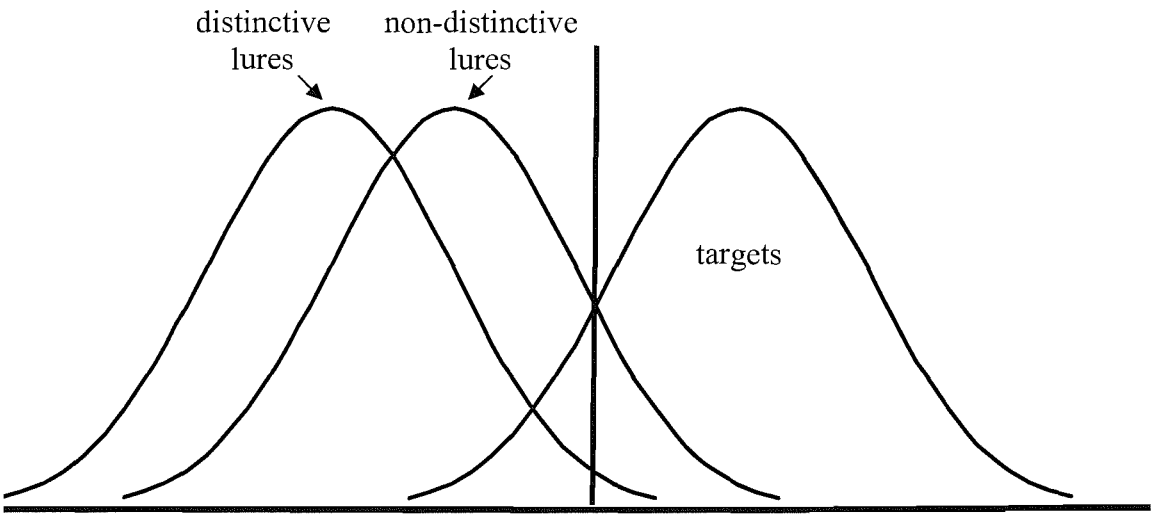
hence the distribution for the former would be further to the right than that for the latter items. Because there is no reason that lures in both conditions should differ in levels of familiarity, the FA suppression found in the distinctive encoding condition is then produced by assuming a more conservative criterion setting here than in the non-distinctive encoding condition. In this case, however, the criterion would have to shift to the *exact* extent such that the hit rate would remain more or less equal across conditions. Although this model is plausible, it inconveniently introduces a new and yet-to-be-explained mechanism, one which allows participants to shift their response criterion to maintain stable hit rates.

Similarly, early likelihood-ratio models, such as that based on Glanzer's ALT (e.g., Glanzer & Adams, 1985, 1990) also has difficulties in reconciling data from research on the distinctiveness heuristic. Through the principle of dispersion (e.g., Hilford et al., 1997), FA rate reduction achieved for the distinctively encoded items should be accompanied by an improvement in the hit rate. The stability of hit rates observed in the number of studies cited above is therefore inconsistent with the predictions of early likelihood-ratio models. It also remains to be seen how newer likelihood-ratio models (e.g., REM; Shiffrin & Steyvers, 1997) can account for the use of the distinctive heuristic in the rejection of lures in conditions where items had been distinctively encoded, and whether simulations of these models can produce the other version of a partial mirror pattern – one with stable hit rates but differences in FA rates across encoding conditions.

In contrast to the criterion-shift and likelihood-ratio models, the multi-process SD model (Tam & Higham, 2006; see also Wixted & Stretch, 2000, 2004) would regard the distinctive heuristic to be a metacognitive process which generates evidence *against* a lure's prior occurrence, and thereby induces a distribution shift (to the left) for the lures on the underlying strength-of-evidence continuum (see Figure 9). The stability of hit rates is accounted for here by assuming a fixed response criterion and a singular distribution for targets studied in the different encoding conditions. Thus, although the experimental manipulation stipulated that items were encoded more distinctively in one condition than the other, recognition performance for targets was actually equivalent in both conditions, and hence the targets form a unitary distribution. In this sense, the implementation of the distinctiveness heuristic

might principally depend on the participant's subjective assessment of item memorability, rather than on the objective measure of item memorability (which is reflected by the item class's hit rate). That is, if the participant *believes* that items were encoded distinctively, then the distinctiveness heuristic would be used to reject lures, regardless of actual hit rate performance. It follows then that this intriguing dissociation between subjective and objective assessment of memorability may underlie the absence of a full mirror effect not only in the distinctiveness heuristic literature, but also in other experimental paradigms described throughout this thesis. In view of this dissociation, it is notable that the multi-process model possesses the required flexibility to accommodate findings of changes in FA rates accompanied by stable hit rates (as in research on the distinctiveness heuristic), or vice versa (i.e., hit rate changes but stable FA rates; e.g., Stretch & Wixted, 1998; Morrell et al., 2002; Hirshman & Arndt, 1997).

Figure 9. A multi-process model for the distinctiveness heuristic. Distinctive lures form a separate distribution which is located lower on the strength-of-evidence dimension than that formed by non-distinctive lures. The criterion is fixed.



8.8 Conclusions and Suggestions for Future Work

The current thesis was motivated by the question concerning the legitimacy of using the hension effect as an empirical illustration of the discrepancy-attribution hypothesis (e.g., Whittlesea & Williams, 1998, 2000). In direct response to this

question, several experimental manipulations were imposed which targeted the perception of discrepancy in the hension effect paradigm (Experiments 1 – 4). These experiments showed that discrepancy did *not* appear to be a critical factor underlying the effect. Indeed, the partial elimination of the hension effect (that between regular and irregular nonwords) when the materials were subsequently revised to reduce inter-item similarity (Experiments 6 – 8), has put doubts on the relevance of the discrepancy-attribution hypothesis in explaining the hension effect. In this way, conclusions drawn from the current thesis are in agreement with the scepticism concerning the role of discrepancy in false recognition (at least in the hension effect paradigm), a sentiment which has also been raised in recent years by other researchers (e.g., Cleary et al., 2005; Reber et al., 2004).

As emphasised at the beginning of the introductory chapter, the objective of the thesis was not to discredit the discrepancy-attribution hypothesis as a whole – the hypothesis may well be a suitable account for other memory and decision-making processes (e.g., Whittlesea & Leboe, 2000, 2003). Based on the findings from the current thesis, however, the hypothesis as a viable account for the hension effect was disputed. The alternative explanation for this phenomenon, as offered by the present thesis, was one centring on memorability-based rejections in recognition. In experiments reported in the latter parts of the thesis, the possibility that item memorability could be used as a basis to correctly reject lures was demonstrated. Importantly, it was shown that even for items whose intrinsic memorability is low (e.g., nonwords), FA suppression for these items was obtained if these items' extrinsic memorability was enhanced through experimental manipulation (i.e., repetition; cf. Experiments 7 & 8). In demonstrating FA suppression by manipulating item strength within list, the experimental findings here went against previous research outcomes (e.g., Stretch & Wixted, 1998; Morrell et al., 2002). It was argued that the novel findings obtained in Experiments 7 and 8 could be attributed to the labelling system devised. Importantly, the effectiveness of this system in producing within-list FA suppression highlighted that in order for memorability cues to be utilised fully, these cues need to be explicit and unambiguous.

One line of future work could therefore focus on the type and saliency of experimental conditions, as well as the effectiveness of memorability cues in effecting

memorability-based rejections. An example of this line of research can be found most recently in Singer and Wixted (2006), who demonstrated that the saliency of the experimental manipulation (extremely long delay between study and test) was a critical factor in predicting whether memorability-based rejections would occur (see section 6.6). In regards to memorability cues, results from the present thesis indicated that the labelling system devised here provided a reliable aid for participants in monitoring the effects of experimental manipulations on item memorability. Future work is needed to ascertain whether this memorability cueing system would be equally effective in inducing FA suppression (or full mirror effects) in paradigms which have previously failed to demonstrate these outcomes in recognition memory.

Another issue which has been emphasised by the present thesis was the comparison between intrinsic and extrinsic memorability (see also Ghetti, 2003; Benjamin & Bawa, 2004). In future research, it will be important to distinguish between memorability from item-based intrinsic factors on the one hand, and from experimental, situational factors on the other. Further, as discussed in previous sections, the rarity of FA suppression for item groups whose memorability had been experimentally enhanced, and the absence of a full mirror effect in the distinctiveness heuristic literature, suggest that participants might have difficulty in accurately assessing item memorability, especially when it had been affected by extrinsic factors. Another line of future research, therefore, can centre on how recognition performance may be positively correlated with the participants' ability to monitor and match their subjective assessment to the objective measure of item memorability (as reflected by hit rates).

To provide a SD account for data from past research and the present thesis, it was proposed here that a multi-process approach could be adopted in modelling not only "old" responses (Wixted & Stretch, 2000, 2004), but also "new" responses (Tam & Higham, 2006). The multi-process model does not necessitate that participants frequently readjust their criterion from item-to-item during the test phase, rather, FA suppression for memorable items is modelled via distribution shifts – more memorable lures form a distribution that is lower on the strength-of-evidence continuum than that formed by less memorable lures. Additionally, given the uncertain involvement of criterion shifts in another FA suppression strategy – the

distinctiveness heuristic (Gallo et al., 2004) – it was proposed here that the multi-process perspective may allow the use of this metacognitive strategy to be modelled in SD terms.

The principal argument put forward in this thesis was that although a large part of research in recognition memory may be focussed on explaining the production of false alarms (e.g., the hension effect; Whittlesea & Williams, 1998), it is equally important to consider the way false alarms are prevented through memorability-based metacognitive mechanisms. That decisional heuristics are integral in everyday recognition judgments is reflected by a growing body of research in this area, which includes a number of recent developmental and clinical studies. For instance, children as young as five have been shown to be aware that factors such as event plausibility and saliency could affect the memorability of past events, and older children (9 year-olds), like adults, could consistently utilise these factors in making correct rejections (Ghetti & Alexander, 2004). At the other end of the developmental spectrum, patients with frontal lobe lesions and Alzheimer's disease, unlike healthy controls, were found to be impaired in making correct rejections for lures (Budson, Dodson, Daffner, & Schacter, 2005; Budson, Dodson, et al., 2005; Kishiyama et al., 2004). In the same vein, the conclusion made in this thesis was that metacognitive strategies can be utilised by participants in counteracting factors (e.g., processing fluency) which could promote false recognition. Finally, in proposing a multi-process perspective in modelling metacognitive processes involved in recognition judgments, the present thesis has offered a SD account which has the potential to encapsulate the myriad of experimental outcomes observed in recognition memory research.

Appendix A

Complete Set of Items from Whittlesea and Williams (2000)

Natural Words (60 Items)

station	fraction	sweater	harmful
daisy	shovel	pleasant	honey
machine	herself	dreadful	idiot
isolate	kitchen	tension	lettuce
familiar	peacock	escape	morphine
detail	primate	stallion	notion
basement	romantic	financial	organ
cripple	planet	clinical	palace
fashion	stable	gamble	proceed
absolute	curtain	stomach	reflect
battery	umpire	delicious	swallow
circle	lesson	predict	slender
disease	ramble	animal	silver
eclipse	tendon	engage	thousand
flower	theory	fortune	volcano

Regular Nonwords (60 Items)

hension	scullet	flemin	binical
vassil	tarrion	trespat	visary
plendon	cament	corbit	sendal
purden	pendon	messel	tamid
framble	blissen	hallid	bandal
fissel	maniper	delicon	pladit
subben	passet	pellis	versal
tummel	arman	sonder	waven
mestic	widicom	flamis	cloral
garder	halbert	loffal	blinden
wipple	potimer	belland	gramen
plander	subble	pramis	bingle
wimber	rogation	lomand	crable
calidon	windon	beckle	hammel
barden	brender	forbal	conder

Irregular Nonwords (60 Items)

stofwus	brectelp	molpeot	gotprilb
hadtace	wastisp	beitgan	blectod
pnafted	predtet	plertsod	cinteaf
gertpris	brifcige	linzted	gepird
meunstah	hoendas	loectad	crinbeelp
coelept	pritdib	kortapif	docytan
notirgin	baxtiod	ufilct	flebscort
blentirp	lekudt	jawidtal	macttap

cadpecht	nectpor	nerbipat	retrork
geppiot	gnotid	plafbegt	munherg
tongiter	frevper	banbige	grifpesel
merfica	tlamnic	cumniste	nododdet
ouetis	pratlond	practeep	cumpreze
glizete	lertisp	rientasle	hilgtreb
pnertap	wicstax	spetighe	bicxawa

Practice Items

Natural Words (2 Items)

dictation	remove

Regular Nonwords (2 Items)

fottle	benible

Irregular Nonwords (2 Items)

pasficht	tokwafis

Appendix B

Standard Instructions Given to Participants Prior to Study and Test Phases

Study Phase

Welcome to the experiment.

The first part of this experiment is a STUDY PHASE, here you will be seeing a list of items, presented to you on the computer screen one at a time.

Your task is to remember these words for a later recognition test. Some of these words will be English words, while others will be non-English words.

Each word will be on the screen for [insert duration].

Try your best to remember each of them!

When you're ready, use the mouse to click on the START button and the words will start appearing on the screen for you to study.

Test Phase

That was the end of the STUDY PHASE. The next part of the experiment is the RECOGNITION TEST. Again, you will see a list of items. Some of these will be English words, while others will be non-English words.

Your task in this RECOGNITION TEST is as follows. [Insert instructions specific to the experiment].

Your task is to decide whether the item is OLD or NEW.

If you think the item is one presented earlier during STUDY PHASE, please PRESS the OLD key on the right. If you think the item is NOT one presented earlier during STUDY PHASE, please PRESS the NEW key on the left.

If you think the item is not one presented to you earlier during STUDY, tell your experimenter that the item is NEW.

[In Experiments 1 – 2, participants were asked to verbally inform the experimenter of the recognition judgment].

If you have any questions, please ask your experimenter now.

When you're ready to start the RECOGNITION TEST. Please click on the START button.

Appendix C

Meaning Labels Used for Natural Words, Regular Nonwords and Irregular Nonwords in Experiment 4

Natural Words

1	station	A place where people wait for trains
2	daisy	A commonly-found plant with many petals
3	machine	A device designed for doing work
4	isolate	To set something apart from others
5	familiar	Something that is frequently encountered
6	detail	An individual part of a whole
7	basement	The level below ground in a house
8	cripple	A person or animal who is disabled
9	fashion	The modern trendy style of dress
10	absolute	Something that is not to be doubted
11	battery	A device used to provide electricity
12	circle	A round geometric figure
13	disease	A pathological condition or illness
14	eclipse	The obscuring of the sun by the moon
15	flower	A plant that blossoms in gardens
16	fraction	A part or portion of a whole
17	shovel	A gardening tool used to move dirt
18	herself	A pronoun that relates to the female
19	kitchen	A place where food is cooked
20	peacock	A bird with colourful tail feathers
21	primate	A class of animals including the apes
22	romantic	Something that expresses love
23	planet	A body that revolves around the sun
24	stable	A place where horses are kept
25	curtain	A material that hangs in a window
26	umpire	Someone who keeps scores in tennis
27	lesson	A session where something is learnt
28	ramble	To move about aimlessly
29	tendon	A band of tissue joining muscle to bone
30	theory	A set of statements to explain a fact
31	sweater	A woollen pullover or jumper
32	pleasant	Something that is nice and enjoyable
33	dreadful	Something that is awful and terrible
34	tension	A condition of strain and stress
35	escape	To break free from confinement
36	stallion	An adult male horse
37	financial	A word describing things relating to money
38	clinical	A word relating to places where patients are treated
39	gamble	To bet on an uncertain outcome
40	stomach	A part of the body where food is digested
41	delicious	Something that is nice to taste
42	predict	To make statements about a future event
43	animal	A living being capable of movement
44	engage	To hold the attention of someone

45	fortune	A large sum of money
46	harmful	Something that can cause injury
47	honey	A sweet fluid gathered by bees
48	idiot	Someone who is foolish or stupid
49	lettuce	A leafy green vegetable used in salads
50	morphine	A powerful drug used to relieve pain
51	notion	A belief, idea, or opinion
52	organ	A musical instrument found in churches
53	palace	A place where kings and queens live
54	proceed	To go forward or move on
55	reflect	To mirror or give back an image
56	swallow	To cause food to pass through the throat
57	slender	Someone who is thin and slim
58	silver	A white shiny metal used in jewellery
59	thousand	A number of one followed by three zeros
60	volcano	A mountain from which lava flows

Regular Nonwords

1	hension	A style of Peruvian pottery
2	vassil	A broad sash worn with a Japanese kimono
3	plendon	A sphere which is flat on the top and bottom
4	purden	A sweet biscuit made from treacle
5	framble	An English Renaissance court dance
6	fissel	To embroider with red-coloured threads
7	subben	To forget someone's name
8	tummel	A traditional dress from Mongolia
9	mestic	The hair on an insect's leg
10	garder	A raffia fabric from Madagaskar
11	wipple	A dance performed at a cotton harvest
12	plander	To blind someone with hot objects
13	wimber	To have extremely small feet
14	calidon	A drink made from ale and dried bread
15	barden	An ancient instrument shaped like a trombone
16	scullet	A style of Peruvian pottery
17	tarrion	A broad sash worn with a Japanese kimono
18	cament	A sphere which is flat on the top and bottom
19	pendon	A sweet biscuit made from treacle
20	blissen	An English Renaissance court dance
21	maniper	To embroider with red-coloured threads
22	passet	To forget someone's name
23	arman	A traditional dress from Mongolia
24	widicom	The hair on an insect's leg
25	halbert	A raffia fabric from Madagaskar
26	potimer	A dance performed at a cotton harvest
27	subble	To blind someone with hot objects
28	rogation	To have extremely small feet
29	windon	A drink made from ale and dried bread
30	brender	An ancient instrument shaped like a trombone
31	flemin	An Australian chocolate and coconut sponge cake
32	trespat	A sequence of melody in Greek music

33	corbit	A cylindrical hat worn by Orthodox priests
34	messel	The hoof of an elephant's foot
35	hallid	A veil worn by Greek women
36	delicon	A marsupial with horns found in South America
37	pellis	A Turkish unit of weight
38	sonder	A believer in two Gods at the same time
39	flamis	A suicide that is disguised as a murder
40	loffal	A unit of measurement in geophysics
41	belland	A type of copper found in regions of China
42	pramis	An Icelandic folk dance performed by a couple
43	lomand	A drink made of fruit juice and white wine
44	beckle	To ferment milk from a horse
45	forbal	To have hiccups continuously for a long time
46	binical	An Australian chocolate and coconut sponge cake
47	visary	A sequence of melody in Greek music
48	sendal	A cylindrical hat worn by Orthodox priests
49	tamid	The hoof of an elephant's foot
50	bandal	A veil worn by Greek women
51	pladit	A marsupial with horns found in South America
52	versal	A Turkish unit of weight
53	waven	A believer in two Gods at the same time
54	cloral	A suicide that is disguised as a murder
55	blinden	A unit of measurement in geophysics
56	gramen	A type of copper found in regions of China
57	bingle	An Icelandic folk dance performed by a couple
58	crable	A drink made of fruit juice and white wine
59	hammel	To ferment milk from a horse
60	conder	To have hiccups continuously for a long time

Irregular Nonwords

1	stofwus	To sweeten a medicine with a syrup
2	hadtace	A form of Arabic poetry
3	pnafted	A camera used to take pictures of the sun
4	gertpris	A collection or set of picture postcards
5	meunstah	A creature with one limb on its head
6	coelept	An ancient custom of eating or feasting outdoors
7	notirgin	An orange which is yet to be ripened
8	blentirp	A Scandinavian bird-like mythical creature
9	cadpecht	A gap or hole between one's teeth
10	geppiot	The art of painting using egg whites
11	tongiter	A pebble with three sides
12	merfica	The hair that is shaved off a monk's head
13	ouetis	A rack, frame or hanger used to dry paper
14	glizete	To whistle through one's teeth
15	pnertap	A bamboo pole used for scaffolding
16	brectelp	To sweeten a medicine with a syrup
17	wastisp	A form of Arabic poetry
18	predtet	A camera used to take pictures of the sun
19	brifcige	A collection or set of picture postcards
20	hoendas	A creature with one limb on its head

21	pritdib	An ancient custom of eating or feasting outdoors
22	baxtiod	An orange which is yet to be ripened
23	lekudt	A Scandinavian bird-like mythical creature
24	nectpor	A gap or hole between one's teeth
25	gnotid	The art of painting using egg whites
26	frevper	A pebble with three sides
27	tlamnic	The hair that is shaved off a monk's head
28	pratlond	A rack, frame or hanger used to dry paper
29	lertisp	To whistle through one's teeth
30	wicstax	A bamboo pole used for scaffolding
31	molpeot	A polka-like Polish dance
32	beitgan	An old style of poetry with unusual rhymes
33	plertsod	An old German title of nobility
34	linzted	A man-shaped sea monster
35	loectad	A patch of cloth inserted in a skirt
36	kortapif	The metallic dust from grinding of metals
37	ufilct	The strip of spacing in lines of printing
38	jawidtal	A fast Hungarian dance
39	nerbipat	A type of lobster found in Samoa
40	plafbegt	An all-night vigil before an Orthodox church feast
41	banbige	To form something into a square-shaped object
42	cumniste	To fire a gun in the air during a duel
43	practeep	An edible sculpture made from pastry
44	rientasle	To fill in mortar joints with small pebbles
45	spetighe	An ancient leather coat worn in battles
46	gotprilb	A polka-like Polish dance
47	blectod	An old style of poetry with unusual rhymes
48	cinteaf	An old German title of nobility
49	gepird	A man-shaped sea monster
50	crinbeelp	A patch of cloth inserted in a skirt
51	docytan	The metallic dust from grinding of metals
52	flebscort	The strip of spacing in lines of printing
53	macttap	A fast Hungarian dance
54	retrork	A type of lobster found in Samoa
55	munherg	An all-night vigil before an Orthodox church feast
56	grifpesel	To form something into a square-shaped object
57	nododdet	To fire a gun in the air during a duel
58	cumpreze	An edible sculpture made from pastry
59	hilgtreb	To fill in mortar joints with small pebbles
60	bicxawa	An ancient leather coat worn in battles

Appendix D

Bigram Frequency

To provide an objective measure of orthographic regularity, the bigram frequencies of all items used in the hension effect paradigm were determined. These frequencies were obtained after consulting the comprehensive count carried out by Solso and Juel (1980), who tabulated the frequencies of all bigrams which appeared in the corpus of one million words collected by Kucera and Francis (1967). Solso and Juel's list of bigram frequencies are sensitive to position and word length in that the frequency of a given bigram varies depending on the position it appears within a word, and the number of letters contained in that word. For example, in 5-letter words, the bigram AF appears 1074 times per million when they occupy the 1st and 2nd positions (e.g., AFTER), but only 180 times per million when they occupy the 3rd and 4th positions (e.g., CRAFT). Similarly, the bigram frequency of AF falls to 240 per million when they appear in the 1st and 2nd positions in 6-letter words (e.g., AFFORD).

Based on the count by Solso and Juel (1980), the average bigram frequency for each item could be obtained simply by computing the mean of all bigram frequencies within that item[1]. For example, the average bigram frequency of the natural word CURTAIN is 824.33 (CU = 373, UR = 1266, RT = 914, TA = 891, AI = 694, and IN = 808). The average bigram frequency for nonwords was calculated in a similar manner. For example, the measure is 1211.67 for the regular nonword HENSION (HE = 556, EN = 1857, NS = 313, SI = 1261, IO = 1387, and ON = 1896), and 474.67 for the irregular nonword STOFWUS (ST = 1704, TO = 315, OF = 55, FW = 18, WU = 0, and US = 756).

Item Category Size of 60 Items. The mean of average bigram frequencies for all 60 items from each item type is shown in Table 23. A one-way ANOVA, carried out to elucidate differences among item groups in terms of bigram frequency, revealed in a significant item main effect $F(2, 179) = 17.53$, $p < .001$, $MSE = 214394.59$, $\eta^2 = .164$. Post-hoc independent-samples t-tests ($\alpha = .0167$) suggested that

[1] Because the item groups differ significantly from each other in terms of item length, the average, rather than the sum, of bigram frequencies for each item here would be more appropriate as a measure of orthographic regularity.

this item main effect arose because the average bigram frequency for irregular nonwords was significantly lower than that of natural words, $t(118) = 4.77, p < .001$, $SE = 393.14$, $\eta^2 = .162$, and regular nonwords, $t(118) = 5.66, p < .001$, $SE = 82.16$, $\eta^2 = .213$. Natural words and regular nonwords did not differ significantly from each other in bigram frequency, $t(118) = .808, p > .42$.

Table 23. The means (and standard deviations) of average bigram frequency, according to item type, and the number of items (60 items or 40 items per item type) contributing to the calculations.

	Average Bigram Frequency			
	60 Items per Type		40 Items per Type	
Natural	924.85	(488.00)	898.73	(498.76)
Regular	996.66	(485.06)	1019.88	(495.77)
Irregular	531.72	(412.01)	628.48	(459.57)

Item Category Size of 40 Items. Table 23 also shows the means of item-average bigram frequency for each item group, after 20 items had been discarded from the original set of 60 items in each category (see section 5.8.3 for more details). In order to ascertain differences in average bigram frequencies among the three item groups (with 40 remaining items in each group), a one-way ANOVA, with item (natural/ regular /irregular) as the between-group factor, was conducted. This analysis was identical to that performed on the original set of items (see above), and it was revealed that for the remaining items, significant inter-group differences in item-average bigram frequencies still existed, $F(2, 117) = 6.83, p < .005$, $MSE = 235248.76$, $\eta^2 = .105$. This effect was driven by the way that item-average bigram frequency was greater for both natural words and regular nonwords, than for irregular nonwords, $t(78) = 2.52, p < .0167$, $SE = 107.23$, $\eta^2 = .075$, and $t(78) = 3.66, p < .001$, $SE = 106.89$, $\eta^2 = .147$ respectively. The bigram frequency measure did not differ significantly between natural words and regular nonwords, $t(78) = 1.09, p > .25$. It should be noted that despite the removal of one third of the items from the original

stimulus set, this analysis on bigram-frequency differences among item groups yielded the identical outcome to the analysis on the original, full set of items (see above). In both analyses, orthographic regularity (measured by bigram frequencies) was similar for natural words and regular nonwords, and both of these item types were orthographically more regular than irregular nonwords.

Appendix E

Item Length

Item Category Size of 60. A one-way ANOVA, with item type (natural/ regular/ irregular) as the independent variable, was conducted on item length. It was found that the item main effect was significant, $F(2, 179) = 31.99, p < .001$, $MSE = .585$, $\eta^2 = .263$. Post-hoc independent-samples t-tests ($\alpha = .0167$) indicated that all three item types differ significantly from each other in length. On average, irregular nonwords are longer than natural words, $t(118) = 3.80, p < .001$, $SE = .153$, $\eta^2 = .109$. In turn, natural words are on average longer than regular nonwords, $t(118) = 3.76, p < .001$, $SE = .142$, $\eta^2 = .107$.

Item Category Size of 40. The same one-way ANOVA, with item type (natural/ regular/ irregular) as the independent variable, was performed on item length after the removal of items high in length or in inter-stimulus similarity. As in the previous analysis, the item main effect was significant, $F(2, 119) = 9.979, p < .001$, $MSE = .482$, $\eta^2 = .144$. This item main effect arose because regular nonwords are still significantly shorter in length than both natural words, $t(78) = 2.813, p < .01$, $SE = .169$, $\eta^2 = .092$, and irregular nonwords, $t(78) = 4.927, p < .001$, $SE = .137$, $\eta^2 = .237$. Unlike the previous analysis, however, the item length of natural words and irregular nonwords was not significantly different from each other, $t(78) = 1.27, p > .20$.

Table 24. The means (and standard deviations) of item length, according to item type and the number of items (60 or 40) in each category.

	60 Items		40 Items	
Natural	6.82	(.93)	6.85	(.83)
Regular	6.28	(.58)	6.37	(.67)
Irregular	7.40	(.74)	7.05	(.55)

Appendix F

Items Used in Experiments 6 – 10

Natural Words (40 Items)

daisy	herself	theory	honey
machine	peacock	sweater	morphine
isolate	primate	pleasant	notion
detail	romantic	dreadful	palace
basement	stable	escape	proceed
cripple	curtain	stallion	reflect
fashion	umpire	predict	swallow
absolute	lesson	engage	slender
eclipse	ramble	fortune	thousand
fraction	tendon	harmful	volcano

Regular Nonwords (40 Items)

hension	scullet	rogation	beckle
purden	tarrion	brender	binical
fissel	cament	flemin	visary
subben	blissen	trespat	sendal
tummel	maniper	corbit	tamid
mestic	passet	delicon	waven
garder	arman	sonder	cloral
plander	widicom	belland	blinden
wimber	halbert	pramis	gramen
calidon	potimer	lomand	crable

Irregular Nonwords (40 Items)

stofwus	tongiter	baxtiod	linzted
hadtace	merfica	lekudt	loectad
pnafted	ouetis	nectpor	ufilct
gertpris	glizete	gnotid	banbige
meunstah	pnertap	frevper	blectod
coelept	brectelp	tlamnic	cinteaf
notirgin	wastisp	lertisp	gepird
blentirp	predtet	wicstax	docytan
cadpecht	hoendas	molpeot	macttap
geppiot	pritdib	beitgan	retrork

Appendix G

Inter-Stimulus Similarity

From the original set of 180 items (60 items from each stimulus category), items that resembled others in terms of orthography and phonology, as well as the majority of items exceeding 7 letters in length, were discarded. The remaining items were 120 in total (40 items from each category). For these 120 items, two indices of inter-stimulus similarity could be calculated – before and after the downsizing of the stimulus pool. Inter-stimulus similarity of an item is defined here as the number of other items in the stimulus pool with which the particular item shares its first and/or last three phonemes. For a particular item, this index was expected to fluctuate depending on the size of the stimulus pool. For example, one of the regular nonwords retained was TUMMEL. Before the stimulus pool was downsized, TUMMEL had a similarity index of 2 because it shares its last three phonemes with two other items – ANIMAL and HAMMEL. After downsizing (where both ANIMAL and HAMMEL were removed), the similarity index of TUMMEL dropped to 0.

A mixed 3 (item: natural/ regular/ irregular) x 2 (pool size: original/ reduced) ANOVA was therefore performed on the inter-stimulus similarity of the 120 items which remained after the stimulus pool size had been reduced (see means in Table 25). As this was an item-analysis, the between-subjects factor was item, and the within-subjects factor was pool size. The two levels of the pool size factor (original and reduced) referred to whether the inter-stimulus similarity index was calculated based on the original (180 items) or the reduced (120 items) pool size. This analysis resulted in a significant main effect of pool size, $F(1, 117) = 34.39, p < .001, MSE = .577, \eta^2 = .227$, reflecting that the inter-stimulus similarity for the 40 items was greater when the pool size was large (original, $M = 1.26$) than when it was small (reduced, $M = .68$). Both the item main effect, as well as the item x pool size interaction, were significant, $F(2, 117) = 7.24, p < .002, MSE = 4.33, \eta^2 = .110$, and $F(2, 117) = 7.09, p < .002, MSE = .577, \eta^2 = .108$ respectively.

Post-hoc t-tests, with Bonferroni adjustment to the alpha ($\alpha = .0167$) showed that the item main effect arose because averaged across pool size, inter-stimulus similarity was higher for regular nonwords ($M = 1.58$) than irregular nonwords ($M = .33$), $t(78) = 4.23, p < .001, SE = .296, \eta^2 = .187$. Natural words' mean similarity

index ($M = 1.01$) was also higher than that for irregular nonwords, but this comparison was marginally significant, $t(78) = 2.41, p < .02, SE = .285, \eta^2 = .069$. However, natural words and regular nonwords were not found to differ significantly on this index, $t(78) = 1.42, p > .15$.

Post-hoc t-tests ($\alpha = .0167$) were also performed as follow-up analyses to the significant item x pool size interaction. These analyses revealed that the interaction arose because the decrease in inter-stimulus similarity, as a result of reduction in pool size, was significant for both regular nonwords, $t(39) = 4.42, p < .001, SE = .226, \eta^2 = .334$, and natural words, $t(39) = 3.44, p < .002, SE = .181, \eta^2 = .233$. However, the same decrease in the similarity index was not significant for irregular nonwords, $t(39) = 2.08, p > .04$.

Table 25. The means (and standard deviations) of inter-stimulus similarity for the final list of 40 items, according to item type and pool size (either original or reduced).

	Original		Reduced	
Natural	1.33	(2.22)	.70	(1.30)
Regular	2.08	(2.36)	1.08	(1.40)
Irregular	.38	(.59)	.28	(.45)

Appendix H

Additional Items Used in Experiments 9 – 10

Natural Words (4 Items)

animal	battery	circle	disease

Regular Nonwords (18 Items)

surdic	dessdom	hevent	tandion
winsial	meckry	plorier	selint
scanser	poisert	canicat	clisper
dellmer	flasand	pulban	
wirbet	bemmet	bramel	

Irregular Nonwords (4 Items)

nerbid	kortapt	widtal	pladege

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering Can Cause Forgetting - Retrieval Dynamics in Long- Term-Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063-1087.

Andrews, S. (1989) Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 802-814.

Andrews, S. (1992) Neighbourhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 234-254.

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds?. *Journal of Experimental Psychology: Human Perception and Performance, 24*(4),1052-1086.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language, 39*, 371-391.

Arndt, J., & Reder, L. M. (2003). The effect of distinctive visual information on false recognition. *Journal of Memory and Language, 48* (1), 1-15.

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing speed and accuracy of word recognition. In S. Kornblum (Ed.), *Fourth international symposium on attention and performance* (pp. 583-611). New York: Academic Press.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 1. Learning, memory & thinking*. San Francisco: Freeman.

Baddeley, A. (1982). Domains of recollection. *Psychological Review, 89*, 708-729.

Baird, J. C., & Noma, E. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

Bartlett, F.C. (1932). *Remembering: An experimental and social study.* Cambridge: Cambridge University Press.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4), 941-947.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31*(2), 297-305.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51*, 159-172.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309-330). Hillsdale, NJ: Erlbaum.

Bransford, J. D. & Franks, J. J. (1971).The abstraction of linguistic ideas. *Cognitive Psychology 2*, 331-350.

Brown, J. (1976). An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition* (pp.1-35). New York: Wiley.

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.

Budson, A. E., Dodson, C. S., Daffner, K. R., & Schacter, D. L. (2005). Metacognition & false recognition in Alzheimer's disease: Further explorations of the distinctiveness heuristic. *Neuropsychology, 19*, 253-258.

Budson, A. E., Dodson, C. S., Vatner, J. M., Daffner, K. R., Black, P. M., & Schacter, D. L. (2005). Metacognition & false recognition in Alzheimer's disease and in patients with frontal lobe lesions. *Neuropsychologia, 19*, 253-258.

Budson, A. E., Droller, D. B. J., Dodson, C. S., Schacter, D. L., Rugg, M. D., Holcomb, P. J., & Daffner, K. R. (2005). Electrophysiological dissociation of picture versus word encoding: Understanding the distinctiveness heuristic as retrieval orientation. *Journal of Cognitive Neuroscience, 17*, 1181-1193.

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*, 231-248.

Chalmers, K. A., & Humphreys, M. S. (1998). Role of generalized and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 610-632.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*, 37-60.

Cleary, A. M., Morris, A. L., & Langley, M. M. (2005). *The impact of inherent stimulus properties on recognition memory: When a novel stimulus adheres to a known structural regularity*. Unpublished manuscript.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review, 108*, 204 - 258.

Coombs, C., Dawes, R. M., & Tversky, A. (1970). *Mathematical Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Criss, A. H., & Shiffrin, R. M. (2004). Interactions between study task, study time, and the low frequency hit rate advantage in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 778-786.

Deffenbacher, K. A., Johanson, J., Vetter, T., & O'Toole, A. J. (2000). The face typicality-recognizability relationship: Encoding or retrieval locus. *Memory & Cognition, 28*(7), 1173-1182.

Dewhurst, S. A., & Anderson, S. J. (1999). Effects of exact and category repetition in true and false recognition memory. *Memory & Cognition, 27*(4), 665-673.

Dewhurst, S. A., Holmes, S. J., Brandt, K. R., & Dean, G. M. (2006). Measuring the speed of the conscious components of recognition memory: Remembering is faster than knowing. *Consciousness and Cognition, 15*, 147-162.

Dobbins, I. G., & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an it em-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1186-1198.

Dobbins, I. G., Kroll, N. E. A., Yonelinas, A. P., & Liu, Q. (1998). Distinctiveness in recognition and free recall: The role of recollection in the rejection of the familiar. *Journal of Memory and Language, 38*, 381-400.

Dodson, C. S., & Schacter, D. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8*(1), 155-161.

Dodson, C. S., & Schacter, D. (2002). When false recognition meets metacognition: The distinctiveness heuristic. *Journal of Memory and Language, 46*, 782-803.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24*, 523-533.

Dunn, J. D. (2004). Remember-know: A matter of confidence. *Psychological Review, 111*(2), 524-542.

Farrell, S., & Lewandowsky, S. (2003). Dissimilar items benefit from phonological similarity in serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), *838-849*.

Foley, M. A., Johnson, M. K. & Raye, C. L. (1983). Age-related changes in confusion between memories for thoughts and memories for speech. *Child Development, 54*, 51-60.

Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 120-128.

Gallo, D. A., Roediger III, H. L., & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review, 8* (3), 579-586.

Gallo, D. A., Weiss, J. A., & Schacter, D. L. (2004). Reducing false recognition with criterial recollection tests: Distinctiveness heuristic versus criterion shifts. *Journal of Memory and Language, 51*, 473-493.

Gardiner, J. M. (1988). Recognition Failures and Free-Recall Failures - Implications for the Relation between Recall and Recognition. *Memory & Cognition, 16*(5), 446-451.

Gardiner, J. M., & Gregg, V. H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review, 4*(4), 474-479.

Gardiner, J. M., & Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition, 19*(6), 617-623.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory, 10*, 83-98.

Ghetti, S. & Alexander, K. W. (2004). "If it happened, I would remember it:" Strategic use of event memorability in the rejection of false events. *Child Development, 75*, 542-561.

Ghetti, S. (2003). Memory for nonoccurrences: The role of metacognition. *Journal of Memory and Language, 48*, 722-739.

Ghetti, S., Qin, J. J., & Goodman, G. S. (2002). False memories in children and adults: Age, distinctiveness, and subjective experience. *Developmental Psychology, 38*, 705-718.

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., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 21-31.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*, 546-567.

Green, D. M., & Swets, J. A. (1966). *Signal detection and psychophysics*. New York: Wiley.

Greene, R. L. (1999). The role of familiarity in recognition. *Psychonomic Bulletin & Review, 6*, 309-312.

Greene, R. L. (2004). Recognition memory for pseudowords. *Journal of Memory and Language, 50*, 259-267.

Greene, R. L., & Thapar, A. (1994). Mirror effect in frequency discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 946-952.

Gregg, V. H., & Gardiner, J. M. (1994). Recognition memory and awareness: A large effect of study-test modalities on "know" responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology, 6*(2), 131-147.

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.

Guttentag, R. E., & Carroll, D. (1997). Recollection-based recognition: Word frequency effects. *Journal of Memory and Language, 37*, 502-516.

Guttentag, R., & Carroll, D. (1998). Memorability judgments for high- and low-frequency words. *Memory & Cognition, 26*(5), 951-958.

Hay, J. F., & Jacoby, L. L. (1996). Separating habit and recollection: Memory slips, process dissociations, and probability matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1323-1335.

Hay, J. F., & Jacoby, L. L. (1999). Separating habit and recollection in young and older adults: Effects of elaborative processing and distinctiveness. *Psychology and Aging, 14*(1), 122-134.

Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1105-1120.

Higham, P. A., & Vokey, J. R. (2000). Judgment heuristics and recognition memory: Prime identification and target-processing fluency. *Memory & Cognition, 28*(4), 574-584.

Higham, P. A., & Vokey, J. R. (2004). Illusory recollection and dual-process models of recognition memory. *The Quarterly Journal of Experimental Psychology A, 57*, 714-744.

Hilford, A., Glanzer, M., & Kim, K. (1997). Encoding, repetition, and the mirror effect in recognition memory: Symmetry in motion. *Memory & Cognition, 25*(5), 593-605.

Hintzman, D. L. (1970). Effects of repetition and exposure duration on memory. *Journal of Experimental Psychology, 83*, 435-444.

Hintzman, D. L. (1976). Repetition and memory. In G. H., Bower (Ed.) *The psychology of learning and motivation: Vol. 10. Advances in theory and research* (pp. 47-91). New York: Academic Press.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*, 1-18.

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.

Hirshman, E., & Arndt, J. (1997). Discriminating alternative conceptions of false recognition: The cases of word concreteness and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(6), 1306-1323.

Hirshman, E., & Hostetter, M. (2000). Using ROC curves to test models of recognition memory: The relationship between presentation duration and slope. *Memory & Cognition, 28*(2), 161-166.

Hirshman, E., & Palij, M. (1992). Rehearsal and the word frequency effect in recognition memory. *Journal of Memory and Language, 31*, 477-487.

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review, 2*, 105-112.

Huppert, F. A., & Piercy, M. (1976). Recognition memory in amnesic patients: Effects of temporal context and familiarity of material. *Cortex, 12*, 3-20.

Israel, L. & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4*, 577-581.

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). Deceiving the elderly: Effects of accessibility bias in cued-recall performance. *Cognitive Neuropsychology, 16*(3-5), 417-436.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306-340.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118*(2), 126-135.

Jacoby, L. L., & Witherspoon, D. (1982). Remembering without awareness. *Canadian Journal of Psychology, 36*(2), 300-324.

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(2), 240-247.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391-422). Hillsdale, NJ: Lawrence Erlbaum.

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56*(3), 326-338.

Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General, 118*(2), 115-125.

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

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67-85.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal of Psychology, 94*, 37-64.

Johnston, W. A., Dark, V. J., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(1), 3-11.

Johnston, W. A., Hawley, K. J., & Elliott, J. M. G. (1991). Contribution of perceptual fluency to recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(2), 210-223.

Jones, C. M., & Heit, E. (1993). An evaluation of the total similarity principle: Effects of similarity on frequency judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 799-812.

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.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kishiyama, M. M., Yonelinas, A. P. (2003). Novelty effects on recollection and familiarity in recognition memory. *Memory & Cognition, 31*, 1045-51.

Kishiyama, M. M., Yonelinas, A. P., & Lazzara, M. M. (2004). The von Restorff effect in amnesia: The contribution of the hippocampal system to novelty-related memory enhancements. *Journal of Cognitive Neuroscience, 16*, 15–23.

Kucera, H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Leboe, J. P., & Whittlesea, B. W. A. (2002). The inferential basis of familiarity and recall: Evidence for a common underlying process. *Journal of Memory and Language, 46*(4), 804-829.

Lloyd, M. E., Westerman, D. L., & Miller, J. M. (2003). The fluency heuristic in recognition memory: the effect of repetition. *Journal of Memory and Language, 48*, 603-614.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, Mass: Harvard University Press.

Luo, C. R. (1993). Enhanced feeling of recognition: Effects of identifying and manipulating test items on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 405-413.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

Maddox, W. T., & Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 539-559.

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 616-630.

Malmberg, K. J., & Nelson, T. O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition, 31*(1), 35-43.

Mandler, G. (1979). Organization and repetition: Organizational principles with special reference to rote learning. In L. G. Nilsson (Ed.), *Perspectives on memory research* (pp. 293-327). Hillsdale, NJ: Erlbaum.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*(3), 252-271.

Mandler, G., Nakamura, Y., & Van Zandt, B. J. S. (1987). Nonspecific effects of exposure on stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 646-648.

Marcel, A. J. (1983). Conscious and unconscious perception: An approach to relations between phenomenal experience and perceptual processes. *Cognitive Psychology, 15,* 238- 300.

McCabe, D. P., Presmanes, A. G., Robertson, C. L., & Smith, A. D. (2004). Item-specific processing reduces false memories. *Psychonomic Bulletin & Review, 11*(6), 1074-1079.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724-760.

McNicol, D. (1972). *A primer of signal detection theory*. London: George Allen & Unwin Ltd.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review, 106*, 398-405.

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.

Morris, C. D., Bransford, J. D. & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behaviour, 16*, 519-533.

Norman, K. A. & Schacter, D. L. (1997). False recognition in young and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition, 25*, 838-848.

Odegard, T. N, & Lampinen, J. M. (2005). Recollection rejection: Gist cuing of verbatim memory. *Memory & Cognition, 33*(8), 1422-1430.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconception about signal detection theory. *Psychonomic Bulletin & Review, 10*(3), 556-569.

Peereman, R., & Content, A. (1995). The Neighbourhood size effect in naming: lexical activation or sublexical correspondences? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 409-421.

Poldrack, R. A., & Logan, G. D. (1998). What is the mechanism for fluency in successive recognition? *Acta Psychologica, 98*, 167-181.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Rajaram, S. (1993). Remembering and knowing - Two means of access to the personal past. *Memory & Cognition, 21*(1), 89-102.

Read, J. D. (1996). From a passing thought to a false memory in 2 minutes: confusing real and illusory events. *Psychonomic Bulletin & Review, 3*, 105-11.

Reber, R., Zimmermann, T. D., & Wurtz, P. (2004). Judgments of duration, figure-ground contrast and size for words and nonwords. *Perception and Psychophysics, 66*, 1105-1114.

Roediger III, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803-814.

Roediger III, H. L., & McDermott, K. B. (1999). False alarms about false memories. *Psychological Review, 106*(2), 406-410.

Roediger III, H. L., & McDermott, K. B. (2000). Distortions of memory. In E. Tulving & F. I. M. Craik (Eds.), *Oxford handbook of memory* (pp. 149-162). London: Oxford University Press.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43*, 67-88.

Ruiz, J. C., Soler, M. J., & Dasi, C. (2004). Study time effects in recognition memory. *Perceptual and Motor Skills, 98*(2), 638-642.

Sahakyan, L., & Delaney, P. F. (2003). Can encoding differences explain the benefits of directed forgetting in the list method paradigm? *Journal of Memory and Language, 48*, 195-206.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional states. *Psychological Review, 69*, 379-399.

Schacter, D. L., Cendan, D. L., Dodson, C. S., & Clifford, E. R. (2001). Retrieval conditions and false recognition: Testing the distinctiveness heuristic. *Psychonomic Bulletin and Review, 8*, 827-833.

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language, 40*, 1-24.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49*, 289-318.

Schacter, D. L., Verfaellie, M., & Anes, M. D. (1997). Illusory memories in amnesic patients: Conceptual and perceptual false recognition. *Neuropsychology, 11*, 331-342.

Schacter, D. L., Verfaellie, M., & Pradere, D. (1996). The neuropsychology of memory illusions: False recall and recognition in amnesic patients. *Journal of Memory and Language, 35*, 319-334.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 267-287.

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., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34*, 125-137.

Singer, M., Gagnon, N., & Richards, E. (2002). Question answering strategy: The effect of mixing test delays. *Canadian Journal of Experimental Psychology, 56*, 28-64.

Smith, R. E., & Hunt, R. R. (1998). Presentation modality affects false memory. *Psychonomic Bulletin & Review, 5*, 710-715.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34-50.

Solso, R. L., & Juel, C. L. (1980). Positional frequency and versatility of bigrams for two- through nine-letter English words. *Behavior Research Methods and Instrumentation, 12*, 297-343.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language, 33*, 203-217.

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.

Tam, H., & Higham, P. (2006). *Memorability and recognition memory: False alarm suppression within list.* Manuscript submitted for publication.

Tulving, E. (1982). Synergistic ecphory in recall and recognition. *Canadian Journal of Psychology, 36*, 130-147.

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

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*, 1-12.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247*(4940), 301-306.

Tversky, A., & Kahneman, D. L. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Tversky, A., & Kahneman, D. L. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*(3), 291-302.

Vokey, J. R., & Read, J. D. (1995). Memorability, familiarity, and categorical structure in the recognition of faces. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition: Explorations in face space.* (pp. 113-137). London: Routledge Ltd.

Watkins, M. J., & Gibson, J. M. (1988). On the relation between perceptual priming and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 477-483.

Watkins, M. J., & Peynircioglu, Z. F. (1990). The revelation effect: When disguising test items induces recognition. *Journal of Experimental Psyhcology: Learning, Memory, and Cognition, 16*, 1012-1020.

Westerman, D. L., & Greene, R. L. (1996). On the generality of the revelation effect. *Journal of Experimental Psyhcology: Learning, Memory, and Cognition, 22*, 1147-1153.

Westerman, D. L., & Greene, R. L. (1998). The revelation that the revelation effect is not due to revelation. *Journal of Experimental Psyhcology: Learning, Memory, and Cognition, 24*, 377-386.

Westerman, D. L., Lloyd, M. E., & Miller, J. K. (2002). The attribution of perceptual fluency in recognition memory: The role of expectation. *Journal of Memory and Language, 47*(4), 607-617.

Westerman, D. L., Miller, J. K., & Lloyd, M. E. (2003). Change in perceptual form attenuates the use of the fluency heuristic in recognition. *Memory & Cognition, 31*(4), 619-629.

Whittlesea, B. W. A, Masson, M. E. J, & Hughes, A. D. (2005). False memory following rapidly presented lists: The element of surprise. *Psychological Research, 69*(5-6), 420-430.

Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(6), 1235-1253.

Whittlesea, B. W. A. (1997). Production, evaluation and preservation of experiences: Constructive processing in remembering and performance tasks. In D. L. Medin (Ed.), *The psychology of learning and motivation: Vol. 37* (pp. 211-264). New York: Academic Press.

Whittlesea, B. W. A. (2002a). False memory and the discrepancy-attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General, 131*(1), 96-115.

Whittlesea, B. W. A. (2002b). Two routes to remembering (and another to remembering not). *Journal of Experimental Psychology: General, 131*(3), 325-348.

Whittlesea, B. W. A. (2003). On the construction of behavior and subjective experience: The production and evaluation of performance. In J. S. Bowers, & C. J. Marsolek (Eds). *Rethinking implicit memory* (pp. 239-260). New York: Oxford University Press.

Whittlesea, B. W. A. (2004). The perception of integrality: Remembering through the validation of expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 891-908.

Whittlesea, B. W. A., & Leboe, J. P. (2000). The heuristic basis of remembering and classification: Fluency, generation, and resemblance. *Journal of Experimental Psychology: General, 129*(1), 84-106.

Whittlesea, B. W. A., & Leboe, J. P. (2003). Two fluency heuristics (and how to tell them apart). *Journal of Memory and Language, 49*(1), 62-79.

Whittlesea, B. W. A., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition, 29*(2), 234-246.

Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica, 98*(2-3), 141-165.

Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 547-565.

Whittlesea, B. W. A., & Williams, L. D. (2001a). The discrepancy-attribution hypothesis: I. The heuristic basis of feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 3-13.

Whittlesea, B. W. A., & Williams, L. D. (2001b). The discrepancy-attribution hypothesis: II. Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 14-33.

Whittlesea, B. W. A., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language, 29*(6), 716-732.

Wickens, T. D., & Hirshman, E. (2000). False memories and statistical decision theory: Comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychological Review, 107*(2), 377-383.

Witherspoon, D., & Allan, L. G. (1985). The effects of a prior presentation on temporal judgments in a perceptual identification task. *Memory & Cognition, 13*, 101-111.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 681-690.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107*(2), 368-376.

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. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341-1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25*, 747-763.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1415-1434.

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General, 130*(3), 361-379.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

Yonelinas, A. P., Dobbins, I. G., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition, 5*, 418-441.