METAMEMORY AND THE WORD FREQUENCY MIRROR EFFECT

by

Geoffrey John Palmer
B.A. (Hons), University of Manitoba, 2003

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

In the Department of Psychology

© Geoffrey Palmer 2007

SIMON FRASER UNIVERSITY

2007

All rights reserved. This work may not be reproduced in whole or in part, by photocopy or other means, without permission of the author.
APPROVAL

Name: Geoffrey Palmer

Degree: Master of Arts (Psychology)

Title of Thesis: Metamemory and the Word Frequency Mirror Effect

Chair: Dr. Mark Blair
        Assistant Professor

Dr. Bruce Whittlesea
        Senior Supervisor
        Professor

Dr. Thomas Spalek
        Supervisor
        Assistant Professor

External Examiner: Dr. Peter Graf
                    Professor
                    Department of Psychology
                    The University of British Columbia

Date Defended: August 16, 2007
Declaration of Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the "Institutional Repository" link of the SFU Library website <www.lib.sfu.ca> at: <http://ir.lib.sfu.ca/handle/1892/112>) and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, BC, Canada

Revised: Summer 2007
STATEMENT OF ETHICS APPROVAL

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

(a) Human research ethics approval from the Simon Fraser University Office of Research Ethics,

or

(b) Advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University;

or has conducted the research

(c) as a co-investigator, in a research project approved in advance,

or

(d) as a member of a course approved in advance for minimal risk human research, by the Office of Research Ethics.

A copy of the approval letter has been filed at the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Bennett Library
Simon Fraser University
Burnaby, BC, Canada
ABSTRACT

The word frequency mirror effect is the observation that high frequency words produce more false alarms and fewer hits than low frequency words in recognition memory tests (Glanzer & Adams, 1985, 1990). Both single and dual process models of recognition memory have attempted to explain this finding. Single process accounts have argued that both hits and false alarms are the result of familiarity (Glanzer and Adams, 1990), whereas dual process accounts have argued that hits and false alarms are dissociated from each other, with hits being the result of both familiarity and recollection (Joordens and Hockley, 2000; Reder et al., 2000; ). In six experiments, the impact of metamemory judgments on the mirror effect was investigated. Results demonstrated that metamemory judgments affected the hit portion of the effect while the false alarm portion of the effect remained consistent across experiments. The dissociation within the word frequency mirror effect provides support for dual process accounts of recognition memory.

Keywords: memory, metamemory, word frequency, recognition memory
ACKNOWLEDGEMENTS

I am very thankful to all those who have provided support and guidance over the past three years. First and foremost, thank you to Bruce Whittlesea for the advice and motivation he has provided. For his insightful and helpful comments throughout the writing process, I would like to thank Tom Spalek. Thank you to Yifat Faran for her valuable input in the design of the experiments. Finally, thank you to Karen Robson for her constructive feedback on all the drafts of the thesis.
TABLE OF CONTENTS

Approval..............................................................................................................................ii
Abstract ............................................................................................................................ iii
Acknowledgements ........................................................................................................... iv
Table of Contents ............................................................................................................... v
List of Tables ...................................................................................................................... vii

Chapter 1: Introduction........................................................................................................ 1
  Recognition memory Models ......................................................................................... 4
    Source of Activation Confusion (SAC) ........................................................................ 7
    Selective Construction and Preservation of Experiences (SCAPE) .............................. 10
    Signal Detection Theory (SDT) .................................................................................. 13
  The Word Frequency Mirror Effect ............................................................................ 16
    Source of Activation Confusion Explanation of the Mirror Effect ............................... 17
    Selective Construction and Preservation of Experiences Explanation of the Mirror Effect .................................................................................................................. 22
    Attention/Likelihood Explanation of the Mirror Effect ................................................. 24

Chapter 2: Testing the models............................................................................................ 30
  Experiment 1: EORs and Self Paced Training ............................................................... 36
    Method ......................................................................................................................... 38
    Results ......................................................................................................................... 39
    Discussion .................................................................................................................... 44
  Experiment 2: Self-Paced Training .............................................................................. 45
    Method ......................................................................................................................... 46
    Results ......................................................................................................................... 47
    Discussion .................................................................................................................... 49
  Experiment 3: Pleasantness Ratings ............................................................................ 50
    Method ......................................................................................................................... 51
    Results ......................................................................................................................... 52
    Discussion .................................................................................................................... 55
  Experiment 4: EOR First ............................................................................................... 57
    Method ......................................................................................................................... 58
    Results ......................................................................................................................... 59
    Discussion .................................................................................................................... 63
  Experiment 5: Variable Context Training .................................................................... 64
    Method ......................................................................................................................... 65
    Results ......................................................................................................................... 66
LIST OF TABLES

Table 1  Experiment 1: P(old), EORs, and Study Times for High and Low Frequency Words.................................................................41
Table 2  Experiment 1: P(old), EORs and Study Times for High and Low Frequency Words Median Split by Study Time..........................42
Table 3  Experiment 1: P(old), EORs and Study Times for High and Low Frequency Words Median Split by EOR........................................44
Table 4  Experiment 2: P(old), EORs, and Study Times for High and Low Frequency Words........................................................................48
Table 5  Experiment 2: P(old), EORs and Study Times for High and Low Frequency Words Median Split by Study Time......................49
Table 6  Experiment 3: P(old), Pleasantness Ratings (PR), and Study Times for High and Low Frequency Words.................................53
Table 7  Experiment 3: P(old), Pleasantness Ratings (PR) and Study Times for High and Low Frequency Words Median Split by Study Time........54
Table 8  Experiment 3: P(old), Pleasantness Ratings (PR) and Study Times for High and Low Frequency Words Median Split by PR........55
Table 9  Experiment 4: P(old), EORs, and Study Times for High and Low Frequency Words........................................................................60
Table 10  Experiment 4: P(old), EORs and Study Times for High and Low Frequency Words Median Split by Study Time.......................61
Table 11  Experiment 4: P(old), EORs and Study Times for High and Low Frequency Words Median Split by EOR......................................63
Table 12  Experiment 5: P(old) for High and Low Frequency Words with Variable and Consistent Training........................................67
Table 13  Experiment 6: P(old), EORs, and Study Times for High and Low Frequency Words........................................................................71
Table 14  Experiment 6: P(old), EORs and Study Times for High and Low Frequency Words Median Split by Study Time.......................72
Table 15  Experiment 6: P(old), EORs and Study Times for High and Low Frequency Words Median Split by EOR......................................74
CHAPTER 1: INTRODUCTION

Did you have cereal for breakfast this morning? Although answering this question is relatively simple and quick, understanding how humans are capable of answering this question is not nearly as clear. For nearly half a century, researchers from several disciplines have attempted to explain how we are able to accurately recognize an item from our past. Although there has been a wealth of research carried out to explain recognition memory, and much has been discovered about recognition memory, there are still several questions and puzzles that remain unanswered.

The question “Did you have cereal for breakfast this morning?” can be contrasted with the question “What did you have for breakfast this morning?” The former question is a recognition question; the latter is a recall question. Although both refer to the same event, the way the event is remembered is different in each case. In a recognition task, people are given an item (for example cereal), and are asked if they encountered that item in a given context (breakfast). In a recall task, people are given a context, and are asked to generate an item or multiple items that they saw in that context. Both tasks are important to study in order to fully understand human memory.

For a typical person, recall is the most common remembering activity. Trying to remember a grocery list, someone’s name, or directions to a friend’s house are examples of recall. It is fairly rare in everyday life to be asked: “Did you see this item in this context?” Why then is it worth studying recognition memory? While this type of question
might not be a common occurrence in everyday life, studies using recognition memory tests have provided researchers with valuable insight into how memory operates.

This is illustrated by examining the memory literature of the 1960's. Until the early 1960s, most memory research was conducted using recall tests. In a recall test, participants are first presented with a list of items (this is generally referred to as the training phase), and then the participants are asked to report as many of those items as they can remember (this is generally referred to as the test phase). Based on the results of these recall experiments, many researchers argued that forgetting was the result of old information being erased when new information is learned (e.g., Melton & Irwin, 1940). This hypothesis has been called the unlearning hypothesis.

However, as noted by the Generate-Recognize model of recall (Anderson & Bower, 1972) participants in the recall test had to first generate possible candidates and then recognize if these items were studied in the training phase of the experiment. As such, it is possible that either, or both, of those processes were the source of the impaired performance. In other words, perhaps the information was still present in memory, but the participant simply could not generate it. To test this idea, Postman, Stark and Fraser (1968) instead utilized a recognition test which would allow the participant to merely recognize whether an item had been presented previously because the generation of the item had been done for them. Postman et al. found that indeed performance was much improved in the recognition test relative to the recall test. They therefore concluded that forgetting is not necessarily the result of erasure of information from memory, rather the problem may be due to the inability to access that information.
Even though recognition memory is not used very much in everyday life, there are situations in which accurate recognition is extremely important. For example, recognition memory has been extensively studied in the context of eyewitness testimony and police line-ups (see Lindsay, 1994 or Wells & Olsen, 2003 for reviews). In these cases, the answer to the question "Did you see this person in this context" can have serious implications. Therefore, it is important to continue to conduct research aimed at understanding the recognition memory task itself, such as determining the conditions that bring about maximal accuracy as well as to use these tasks to understand memory in general.

In a lab setting, a typical recognition test consists of a training or study phase as well as a test phase. In the training phase, participants are shown a series of items (for example, words, pictures, phrases, etc.). Then, in a test phase, the participants are randomly shown both items they saw during training ("old" items) and items that they did not see ("new" items). Participants are then asked: "Did you see this item in the training phase?" When analyzing the data in a simple recognition test, there are two measures that researchers typically consider. The first is correct detections, or hits. A hit is when participants accurately claim to have seen an item that was presented in the training phase. The second measure is incorrect detections, or false alarms. A false alarm is when participants mistakenly claim to have seen an item that was not presented in the training phase.

False alarms are important to study because these inaccurate memories can offer clues as to how accurate remembering occurs. False memories can reveal the heuristics and biases that the mind uses to reconstruct the past. There is a vast literature studying
false memories (see Hyman and Loftus, 2002 for a review), and recognition memory tests are more likely than recall tests to yield these false memory effects.

While many researchers have used recognition memory tests to demonstrate various effects that are important for theory construction, other researchers have conducted recognition memory tests in order to develop models that explain how recognition memory operates. An understanding of these models is crucial for one to be able to effectively evaluate the results of a recognition memory experiment.

**Recognition memory Models**

In the recognition models that have been proposed, researchers have focused on the relative contributions of two subjective experiences. One of these is the simple feeling that something feels highly familiar but the person does not consciously remember having seen the item in the specific context that is being studied. Henceforth this experience will be referred to as the person’s feeling of “familiarity” regarding that item. A common example of this type of scenario in everyday life would be when you see someone that you are sure that you have seen before, but are unable to pinpoint the exact context in which you had previously seen that person. The other experience is one in which the person does have the conscious memory of having experienced the item in the particular context that is being studied. This experience will be referred to as the person’s “recollection” of the item. An example of this particular scenario would be when you see someone that you remember and you are also able to remember that he was a member of your old bowling team (i.e., the original context in which you encountered that person).
In order to understand recognition memory, the following questions must be answered. First, what are the psychological processes behind familiarity and recollection? Are the two processes fundamentally different from each other? And finally, how is information represented in memory?

Perhaps the first scientific attempt to answer these questions was made by Atkinson and Juola in 1973. They argued that when a participant encounters a stimulus in a remembering task, they will experience a certain level of familiarity. If the familiarity level is very low, then participants will quickly claim the word to be “new”. If the familiarity level is very high, then participants will quickly respond with an “old” response. If the familiarity level is somewhere between the high and low thresholds, then the participant will engage in a search to determine the context in which the item was originally encountered. If the context encountered at study is retrieved, then the participant will respond with an “old” response. If the study context is not retrieved, then the participant will respond with a “new” response. Atkinson and Juola argued that the ease of processing of an item will determine where it falls on the continuum of familiarity. An item that is processed easily produces a high level of familiarity. Recollection, on the other hand, requires that the participant brings to mind the specific memory trace as a result of engaging in a memory search for the context, which would only occur in situations where the item had a moderate familiarity level.

Atkinson and Juola (1973) argued that recollection and familiarity are based on two qualitatively different processes. Although the specific aspects of the models differ, many of the current theories regarding recognition memory are based on the assumption that recollection and familiarity are different processes (Diana, Reder, Arndt & Park,
2006; Joordens & Hockley, 2000; Leboe & Whittlesea, 2002). As a result, these theories are referred to as dual process models of recognition.

Mandler (1980) added to the model proposed by Atkinson and Juola (1973) by defining the specific retrieval processes that occur when experiencing either familiarity or recollection. Mandler argued that the features of a stimulus must be integrated in order for the observer to be able to perceive and identify that stimulus. This is called intra-item integration. When someone attempts to identify a stimulus they have previously encountered, the features of that stimulus will be integrated with each other faster and with greater ease than will be the case for novel stimuli because there is a representation for that item already in stored in memory. The faster that feature-integration occurs, then the greater will be the feeling of familiarity.

Recollection, on the other hand, requires that a search of memory be conducted and that the original context is retrieved. To be successful, this search of memory requires that the participant integrates the item into a specific context. This process is called extra-item integration. If no context is found that the stimulus can be integrated with, the stimulus in question will be deemed to be a “new” item.

While these theories attempted to explain the differences between recollection and familiarity, they did not provide insight into how information is represented in memory. Several current theories attempt to address all three questions highlighted earlier. The three models that will be discussed within this thesis are the Source of Activation Confusion (SAC) model, the Selective Construction and Preservation of Experiences (SCAPE) framework, and Signal Detection Theory (SDT).
Source of Activation Confusion (SAC)

The SAC model proposed by Reder and colleagues (Diana, Reder, Arndt & Park, 2006; Reder, Angstadt, Cary, Erickson & Ayers, 2002; Reder et al. 2000) is consistent with a separate systems approach to memory (e.g., Tulving, 1985). According to the separate systems approach, memory is comprised of several systems that function independently of each other. Each system has separate mechanisms for encoding, storage and retrieval. Two of the systems proposed by Tulving are the episodic system and the semantic systems. The episodic system represents specific experiences whereas the semantic system represents general concepts. One can think of a concept as being the collective experience a person has had with a given stimulus. For example, all of your experiences with chairs would be combined together in order to form your concept of a chair. In other words, your representation for “chair” in semantic memory is the average of all the chairs that you have ever experienced. It is argued that the episodic system is engaged when one performs a remembering task and the semantic system is engaged when one performs any non-remembering task which makes use of past experiences (e.g., identification and classification). Several studies and effects have been taken as evidence that there are several separate memory systems (Jacoby & Dallas, 1981; Tulving, 1985; Tulving, Schacter & Stark, 1983).

The SAC model asserts that information is represented in memory by a network of interconnected nodes. A node can represent either an event or a concept. Concept nodes represent the collective experience that a person has with a given stimulus. Event nodes, on the other hand, represent specific experiences. For example, your current experience of reading this thesis will be represented as an event node. These nodes operate on the principles of activation and inhibition.
When a stimulus is encountered in the world, the corresponding concept node becomes activated and a new event node is formed. For example, if the word TABLE is encountered in the training phase of a memory test, then the concept node for TABLE will become activated. Furthermore, an event node for that particular experience with the word TABLE will be formed. This event node is linked with the concept node and is also attached to specific context nodes to help distinguish it from other event nodes related to that concept. Concept nodes are also linked with each other so that when a concept becomes activated, that activation spreads to all related concept nodes. This spreading activation between concept nodes is thought to be the underlying mechanism behind the semantic priming effect, which is the observation that responding to a word on a non-remembering task is faster if it has been preceded by a related word (Anderson & Bower, 1972; 1973).

Non-remembering tasks, such as identification and classification, are done on the basis of activation of the conceptual nodes. For example, when the word TABLE is presented, the concept node of TABLE is activated and then spreads through the network to other related concept nodes (for example, in this case, the concept node for CHAIR). In this case, because the TABLE node has become activated, it can be identified. Furthermore, the concept for CHAIR is primed and will be identified or classified faster if it is presented immediately after TABLE than some other word not semantically related to TABLE.

Performance on a remembering task, however, involves the additional activation of the event nodes and the contextual nodes. In a recognition test, participants cannot remember that they had seen the word TABLE in a training phase on the basis of
activation of the concept node. This is because the semantic nodal network is comprised of the participant's entire history of encountering stimuli in the world, not just the stimuli presented in the training phase of the experiment. As such, without the event and contextual nodes, it would be impossible to determine if words presented in a test phase were becoming activated because they had been presented in the training phase of the experiment, or at some other point in the past.

Instead, it is hypothesized that in addition to the activation of the conceptual node, the contextual nodes in the network will also be activated based on the similarity to the existing context, and that this activation will spread to the event nodes to which they are attached. If an event node receives activation from a concept node and a context node, then the pooled activation will likely exceed some threshold, and allow the participants to make an "old" response.

One can think of this convergence of activation from conceptual and contextual nodes as akin to searching memory using cues to zero in on the memory trace. To illustrate this process by way of example, let us say that you are trying to remember a specific instance of encountering your sister on a bus. Thus, you will have the cues of your sister (conceptual node), and a bus (contextual node). Although you may have had several experiences with your sister (i.e., several event nodes that are connected to the conceptual node), and several experiences with buses (i.e., several event nodes connected to the contextual node), the specific experience that you had involving your sister and a bus (i.e., the event node) will become activated because it is the only one that is related to all the cues given to initiate the search of memory. If you have several experiences with
your sister and buses, then more contextual cues may be needed to activate one specific event node above the rest.

The SAC model has both familiarity and recollection process built into the model. When a conceptual node reaches a certain level of activation in the context of a remembering test, the participant will experience a feeling of familiarity. If an event node reaches a certain threshold of activation in the context of a remembering task, the participant will experience recollection.

Finally, any model of recognition memory performance must explain why false alarms occur. SAC argues that because items that are falsely called “old” have no event nodes that are linked to a contextual node for the training phase, those false alarms must be made on the basis of familiarity rather than recollection. SAC contends that certain items are more familiar to subjects before the experiment even begins, and that the resting activation levels for those familiar concepts are higher than for more unfamiliar items. Thus, a familiar item encountered at test, may result in a conceptual activation level that is so high that the cognitive system makes the decision based on the familiarity level alone, and does not wait for the slower recollective process.

Selective Construction and Preservation of Experiences (SCAPE)

According to the SCAPE model (Whittlesea, 1997) memory keeps a record of processing experiences. These processing experiences consist of an interaction between a stimulus (what is being paid attention to), a task (what is being done with the stimulus), memory traces of that stimulus (previous experiences with the stimulus), and a context (the environment in which experiences with the stimulus take place). This interaction between the stimulus, the task, the memory trace, and the context is known as the
stimulus complex. SCAPE argues that people do not hold a memory record of the
stimulus itself; rather, memory is a record of what they did with the stimulus in a given
context. Unlike semantic memory, memories are not organized in any particular fashion.
Memories are reconstructed when the proper cues, either external or internal, are
presented. These cues can be related to the context, the stimulus, or the task.

The idea of the stimulus complex can be illustrated using an example of
functional fixedness. Say that you have to pound a nail into a piece of wood, but you do
not have a hammer. All you have is an axe, a nail and a piece of wood. In this situation,
the axe (the stimulus) is likely to have been used in the past (memory trace) only to split
(task) pieces of wood (context). As such, there is no memory of using the reverse side of
an axe to pound the nail. As a result, the idea to use the axe to pound the nail might not
come to mind.

SCAPE contends that there are two primary functions of mind: a productive
function and an evaluative function. The productive function is responsible for processing
the exterior world. The evaluative function is responsible for assessing what has been
produced by the productive function. Production is thought to be controlled by the
principle of transfer appropriate processing (the idea that success in a task depends on the
availability of resources that are appropriate to that task) whereas the evaluative function
is thought to be controlled by inference and attribution. The cognitive system responds to
the outside world on the basis of both the productive and the evaluative functions.

Both the productive and evaluative functions are thought to control remembering.
For example, in a standard recognition task, a word might come up in the test phase that
is identified very quickly. The identification is the result of the productive function. The
evaluative function assesses the processing of the stimulus and tries to attribute a reason as to why in this case the item was processed so quickly. The evaluative function might attribute the quick processing to having seen the item previously, and if so, the participant will call the item "old". On the other hand, the evaluative function might attribute the quick processing to a source outside the experimental context, and in this case, the participant will call the item "new".

Several studies have suggested that the feeling of familiarity is the result of an attribution about current performance to having experienced that stimulus in the past. According to these studies, when an item is processed fluently (compared to other items), an attribution is made about the processing fluency. In the context of a memory experiment, this processing fluency is often attributed to having seen the item in the training phase of the experiment (Higham & Vokey, 2000; Lindsay & Kelly, 1996; Masson & MacLeod, 1996; Whittlesea, 1993; Jacoby & Dallas, 1981). This attribution to the past is what causes the feeling of familiarity. Other researchers have argued that the feeling of familiarity is only caused by a stimulus that has been processed with surprising (i.e., quicker than expected) fluency (Whittlesea & Williams, 1998, 2000, 2001a, 2001b). Whichever is the case, there is agreement among these researchers that the feeling of familiarity is the result of attributing one's current performance to a past event.

Similar to the SAC model of recognition, the SCAPE framework argues that an item can be called "old" on the basis of familiarity or recollection. Where SCAPE deviates from many other models is in its explanation of the underlying cause of recollection. While it is widely accepted that the feeling of familiarity is caused by an attributive process, SCAPE argues that recollection is also the result of an attributive
process. Unlike familiarity, the attribution is made about how well a given item fits in a given context. If the fit between the item and the supposed context is high, the result is a subjective feeling of recollection (Leboe & Whittlesea, 2002). The difference between recollection and familiarity is not that familiarity is the result of an attributive process and recollection is not; the difference is that familiarity is an attributive process about the item itself whereas recollection is an attributive process about the item and the context.

SCAPE explains false alarms through the evaluative function of mind. Because evaluations are made on the basis of inference and attribution, there can be systematic errors. For example, usually when an item is processed with surprising fluency in a recognition test, it is because that item was seen in a training phase. In that case, the surprising fluency is correctly attributed to having seen the item in the training phase. However, it is possible that the source of the surprising fluency comes from a non-experimental context. For example, Whittlesea and Williams (1998) demonstrated that the nonword HENSION is processed fluently despite the fact that it is not a word. If this item is presented in the test phase of a recognition experiment, the participant might misattribute this fluency of processing to the item being old. Often the evaluative function will make the correct attributions, but because the decisions are based on attributions, there can still be errors.

Signal Detection Theory (SDT)

Unlike most models of recognition memory, SDT asserts that recognition judgments are made on the basis of familiarity alone (see MacMillan & Creelman, 1991 for a review). For this reason, it is considered a single-process model of recognition.
The model asserts that before an experiment begins, the items to be used in the experiment have different levels of pre-experimental familiarity. According to SDT, the strength of this pre-experimental familiarity is a source of noise for the recognition decision. When items are encountered during the training phase the familiarity arising from that presentation will add to the pre-experimental familiarity. Therefore, the average familiarity level for items encountered in the training phase will be higher than the average familiarity level for words not encountered in the training phase. The added familiarity that an item receives during the training phase of an experiment is called the signal. After the training phase, there will be two familiarity distributions: one for the words presented in the training phase (signal + noise), and one for the not-presented words (noise). Successful performance on a recognition test depends on whether the added familiarity from the presentation of the item at training (i.e., the signal) is enough to allow one to distinguish between the presented (signal + pre-experimental familiarity) and the not-presented (pre-experimental familiarity) items. In other words, during the test phase of a recognition test, the participant has to try and determine whether the experienced familiarity of that item is more likely caused by the pre-experimental experiences with that item or whether it has added familiarity because of experimental experience.

There are two main variables to consider when explaining performance on a recognition test using SDT. The first is the discriminability (d') between the noise distribution and the signal-plus-noise distribution. In other words, it is the difference in the average familiarity between items that have been studied and items that have not been studied. As can be imagined, there can be huge differences in the d' between experiments.
and between conditions within an experiment. For example, in one experiment, the training phase might consist of reading a list of words very rapidly. In this case, because participants are prevented from elaborately processing the items, there might be only a small difference between the familiarity of the training items and the new items. When the $d'$ is low the two distributions will overlap a great deal so it will be difficult to distinguish from which distribution an item comes from. On the other hand, a second experiment might have a training phase that encourages extensive visual imagery of each item on the training list. In this case, this might result in a huge difference between the familiarity of the training items and the new items. In this case the $d'$ would be high, and there would be very little overlap between the two familiarity distributions. This would cause the items from the two distributions to be relatively easy to distinguish from each other.

The other variable to consider is the criterion that the participants adopt when making the old/new judgments. Some participants may adopt a fairly strict (conservative) criterion for claiming a word to be old; only claiming a word to be old if the item invokes a fairly strong feeling of familiarity. Other participants may adopt a more relaxed (liberal) criterion, whereby the participant will claim a word to be old even if it is only moderately familiar. As participants adopt more conservative criteria, one would expect the participant to make fewer false alarms, but to also miss more of the old words. On the other hand, as participants adopt more liberal criteria, one would expect that the participant to miss fewer of the old words, but would make more false alarms. The decision criteria can be affected by a number of factors such as the personality type of the participant, and the experimental instructions given to the participant.
As mentioned before, SDT explains that false alarms depend on the d' for that particular experimental manipulation, as well as by the decision criteria adopted by the participants. Smaller d' and more liberal decision criteria will both lead to more false alarms.

SDT does not attempt to explain how information is represented in memory, nor does it attempt to explain the underlying process behind familiarity. It does not matter if familiarity is the result of nodal activation, or an attributive process, SDT explains recognition memory in the same manner. What distinguishes SDT from other models is that it explains recognition through familiarity alone, without having to use recollection.

**The Word Frequency Mirror Effect**

The validity of any theory rests on its ability to explain data that has already been collected, and to make predictions about new data. A robust finding in the recognition memory literature is that stimuli at one end of a continuum produce more hits and fewer false alarms than do stimuli at the other end of the continuum. This pattern of data is referred to as a mirror effect. Mirror effects have been found for several dimensions including word frequency (high vs. low frequency words), concreteness (concrete vs. abstract words), meaningfulness (meaningful vs. non meaningful words), and pictures versus words (Glanzer & Adams, 1985, 1990). Mirror effects can be quite puzzling to researchers trying to develop theories of recognition memory. This is because one class of stimuli (e.g. low frequency words) can simultaneously produce more claims of “old” when the item has been studied, and less claims of “old” when the item has not been previously studied. Clearly a strict strength based model of memory cannot account for this pattern of data.
Although the mirror effect can be found using the different types of stimuli mentioned above, this thesis will focus solely on the word frequency mirror effect. The word frequency mirror effect is the finding that relative to high frequency words, low frequency words are responded to more accurately regardless of whether the item is old or new (i.e., higher hit rate and lower false alarms). While the other mirror effects are interesting and important to study, the word frequency mirror effect is the most robust and most studied mirror effect.

Each of the three theories described above provide explanations of the mirror effect. The aim of this thesis will not be to determine which of these theories are correct and which are incorrect. Rather, the aim of the thesis will be to conduct several studies on the mirror effect and then to determine how each of the theories might account for the data that was found. It is anticipated that while some theories might be able to account for the data with little or no adjustment, other theories may require more substantial adjustments.

**Source of Activation Confusion Explanation of the Mirror Effect**

The SAC model explains the mirror effect by theorizing that the activation levels produced in the event and concept nodes by the presentation of new and old words will be different depending on whether the item is a low or a high frequency word. Given that each encounter with a concept will link another event node to that concept node, and that by definition people have more encounters with high frequency words than with low frequency words, Reder et al. (2000) argued that there would be more event nodes linked with the concept nodes of high frequency words than there are for low frequency words. The number of related event nodes linked to a concept is referred to as the *fan* of that
concept (Reder & Anderson, 1980). Thus, high frequency words have a greater fan than low frequency words. Furthermore, Reder et al. argued that there are also differences in the resting activation levels of the concept nodes of high and low frequency words. Because high frequency words are encountered more often, the resting activation level of a high frequency concept node is higher than that of a low frequency word.

As described above, because the SAC model is a dual process account of recognition memory, an old response can be made on the basis of recollection or familiarity. Familiarity is based on the activation of the concept nodes, whereas recollection occurs when the event node becomes activated.

When participants make an "old" response in a recognition test, they will either make a false alarm or a hit. False alarms must be made on the basis of familiarity because there was no new event node formed in the context of the experiment. Familiarity is based on the activation of the concept node, and high frequency words have higher resting levels of activation. Thus, it follows that it is more likely that the activation of a high frequency concept node will reach a threshold that will cause it to feel familiar when compared to a low frequency concept node. As such, participants will be more likely to claim that a new high frequency word is an old item, than they would be for a new low frequency word.

Hits, on the other hand, can be caused by either familiarity or recollection. When the word is encountered in the test phase of an experiment, activation spreads to all related event nodes. However, if participants were using only that test word as a cue, then there would be no way to determine which linked event node was the correct one because they would all have equal activation. There are, however, other contextual cues available
that the participant uses. The cues of “an hour ago” and “on this computer screen” can be used, reducing the number of event nodes that approach the recollection threshold. When activation of a particular event node reaches a certain threshold, the event is recollected. However, because the high frequency words have a greater fan than low frequency words, there is less activation that goes to each individual event node through the cue word and thus it is less likely that the event node formed during the training phase of the experiment will reach the required threshold of activation to elicit a feeling of recollection. Therefore, it is more likely that the low frequency words will be recollected because they have a lower fan than the high frequency words. The recollection advantage for low frequency words outweighs the familiarity advantage for the high frequency words and as a result, there are more hits for the low frequency words.

To test their model, Reder et al. (2000) used the Remember/Know paradigm developed by Tulving (1985). Using this paradigm, when participants claim they have seen a word before, they must say whether they can retrieve the original context of the word (Remember) or if they just have a feeling that they saw it before (Know). Tulving argued that when making a “Remember” judgment, participants are accessing the actual episode, and when a “Know” judgment is made, participants are basing their decision on the activation level of the concept node. Because there is no time and place information associated with the concept node, there cannot be any feeling of remembering. Instead, the participant will simply have a feeling that they encountered the item in the training phase.

Reder et al. (2000) argued that the low frequency words are more likely to be called old because their episodic node is more likely to be retrieved. As such the low
frequency words should be more likely to elicit a “Remember” response from participants. On the other hand, high frequency words are more likely to be called old because their concept node becomes activated, thus, the high frequency words should be more likely to elicit a “Know” response from participants. These were the results that were obtained by Reder and colleagues. Low frequency words were more associated with “Remember” responses, whereas the high frequency words were more associated with “Know” responses.

Reder, Angstadt, Cary, Erikson, and Ayers (2002) also tested the SAC model using pseudo-words. In their experiment, participants learned 80 pseudo-words over twenty sessions that lasted five weeks. In each session, participants were required to study three lists of pseudo-words (referred to as familiarization lists). Participants could only complete one session per day. The study sessions were further grouped into acquisition cycles. Each acquisition cycle consisted of two study sessions (and therefore six familiarization lists). The pseudo-words were separated into four conditions: low frequency, low fan; medium frequency, low fan; medium frequency, high fan; and high frequency, high fan. The frequency of the pseudo-words was defined as the number of times the participants saw each pseudo-word per acquisition cycle. Recall the fan of an item is the number of different contexts that a particular concept is encountered in. The fan of the pseudo-words in this experiment was defined as the number of different lists the pseudo-words were presented on in each acquisition cycle. If a word appeared on one list six times, it had a fan of just one because it had just one context. Words that appeared once on six separate lists had a fan of six because they appeared in six different contexts.
To ensure that participants differentiated the context of the sessions, different fonts and screen colours were used for each session.

The twenty low frequency pseudo-words were studied once per acquisition cycle. By definition, because they could only appear on one list per cycle, all the low frequency words were low fan. The forty medium frequency pseudo-words were studied six times per acquisition cycle. The medium frequency words were divided into high and low fan conditions. In the high fan condition, the pseudo-words were presented in all six familiarization lists per cycle. In the low fan condition, the pseudo-words were presented six times on the same familiarization list. Finally, twenty high frequency words were presented thirty six times per acquisition cycle. They were presented six times on all six lists, and therefore had a high fan. It was impossible to have a low frequency high fan condition because in order for the fan to be high it had to be presented at least six times which would cause the pseudo-word to be medium frequency. Furthermore, there was no high frequency, low fan condition which would require one word to be studied thirty six times on one list.

Recognition tests were given to participants after several sessions. The recognition tests involved a study phase in which some of the pseudo-words were presented followed by a recognition test. Participants had to distinguish if they encountered that pseudo-word in the study phase of that particular recognition test, not simply from the familiarization sessions. All words presented on the test phase of the recognition test had been presented during the familiarization sessions. On the recognition tests that occurred after the later learning cycles, the standard mirror effect was observed: low frequency pseudo-words had more hits and fewer false alarms than
high frequency pseudo-words. The Remember/Know procedure was also employed in this study, and once again, low frequency pseudo-words were more associated with “Remember” responses and high frequency pseudo-words were more associated with “Know” responses. The SAC model also predicted that pseudo-words with a greater fan would be less recollected because less activation would be spread to the related event nodes minimizing the amount of activation that spread to any specific event node. Looking at the medium frequency pseudo-words, it was found that pseudo-words with a high contextual fan were less likely to elicit a “Remember” response than words with a low contextual fan.

These two experiments were taken as evidence by Reder and colleagues (2000, 2002) that both familiarity and recollection are involved in recognition memory. Furthermore, they argued that information is represented in memory by concept, context and event nodes and information is remembered through activation of those nodes.

Selective Construction and Preservation of Experiences Explanation of the Mirror Effect

Like the SAC model, SCAPE (Whittlesea, 1997) argues that false alarms are made, for the most part, on the basis of familiarity and that hits are made on the basis of recollection and familiarity. The difference between the two models is the underlying cause of familiarity and recollection. Reder and colleagues (2000, 2002) argued that remembering is the result of activation of event and concept nodes. In contrast, SCAPE argues that remembering is a constructive process based on inference and attribution.

As stated above, the SCAPE framework argues that processing fluency causes a feeling of familiarity. When an object is processed quickly (or fluently) an attribution is
made about the source of that fluency. In the context of a recognition experiment, the fluency is attributed to having seen the item in the training phase of the experiment. By definition people have more experiences with high frequency words, and therefore the high frequency words will be processed more fluently than the low frequency words. Because they are processed more fluently, it is more likely that the high frequency words will reach the level of fluency that will cause the participant to experience a feeling of familiarity and to call that item an old item. Because false alarms are made, for the most part, on the basis of familiarity, there are more false alarms for high frequency words than for low frequency words.

Correct detections, or hits, can be made on the basis of familiarity or recollection. According to the SCAPE framework, recollection occurs when an attribution is made about how well an item fits into a particular context. Because low frequency words are more distinctive, in the training phase of an experiment low frequency words will be better integrated with the experimental context. Then, in the test phase, there will be a greater fit between the distinctive low frequency words and the context of the experiment than between the high frequency words and the experimental context, leading to more hits for the low frequency words.

Joordens and Hockley (2000) conducted a series of experiments that provided some support for this model. Like Reder et al. (2000), they used the Remember/Know procedure and found that low frequency words were more associated with “Remember” responses and high frequency words are more associated with “Know” responses. Furthermore, Joordens and Hockley also manipulated several factors that would affect the probability of recollecting words. For example, in one series of experiments, they
manipulated the delay between the training and test phase of the experiment. They argued that this manipulation would negatively affect recollection but not familiarity. If this was the case then the hit portion of the mirror effect should be affected while the false alarm portion should remain constant. Joordens and Hockley found that any manipulation that reduced the probability of recollecting of the word eliminated the low frequency hit advantage while leaving the false alarm pattern intact. They argued that this dissociation provided evidence in support of the position that the hit portion of the mirror effect was due to recollection and the false alarm portion was due to familiarity.

While this series of experiments does not provide support for the SCAPE account over the SAC account, it does provide support for the idea that recognition is based on two separate processes, namely familiarity and recollection.

**Attention/Likelihood Explanation of the Mirror Effect**

Glanzer and Adams (1990) proposed the Attention/Likelihood model to account for the word frequency mirror effect. This model uses the assertions of SDT. According to this model, familiarity alone can account for the mirror effect.

The Attention/Likelihood model states that stimuli are simply sets of marked and unmarked features. When a stimulus is encountered in the world, some of those features become marked. Thus, all stimuli have a certain number of marked features before an experiment begins. These pre-experimental marked features can be considered noise. In the experimental setting, when a stimulus is encountered, more features become marked. These marked features can be considered the signal that participants are trying to detect in the test phase of the experiment. It is assumed that low frequency words are given more attention in the training phase because of their distinctive properties and thus have
more features marked. In the test phase of an experiment, the participant sees an item and
determines the proportion of marked features that are present for that stimulus.
Furthermore, the participant determines the proportion of marked features expected for an
old item and a new item of that type. From this information a likelihood ratio is computed
that represents the perceived likelihood that the item is old, and a decision is made based
on the likelihood ratio. If the likelihood ratio exceeds a preset criterion, the participant
will claim that item to be an old item.

The low frequency advantage for hits occurs because of the number of marked
features that these items receive in the training phase of the experiment. The low
frequency words are given more attention and thus have more marked features. In the test
phase, because there are more marked features for the low frequency words, it is more
probable that those words will be called old. However, the Attention/Likelihood model
asserts that all new items have the same number of marked features. Recall that the
likelihood ratios are computed based on the expected proportion of marked features for
that type of word. This means that two items can have the same number of marked
features, but different likelihood ratios will be computed depending on the frequency of
the item. Because the proportion of features marked in the training phase of the
experiment is smaller for the high frequency words, there is more overlap between the
familiarity distribution of high frequency old words and the familiarity distribution of
high frequency new words than the distributions of low frequency old and new words.
This greater overlap will not only cause more old items to be called new, but it will also
cause more new items to be called old. Thus, there will be more false alarms for the high
frequency words than for the low frequency words.
While Reder et al. (2000) and Joordens and Hockley (2000) used dissociations to demonstrate different processes underlying the hits and the false alarms of the mirror effect, Glanzer, Adams, and Iverson (1991) considered a lack of a dissociation to be evidence that the hits and false alarms were controlled by the same process. In their experiment, the test phase consisted of a forced choice recognition test rather than a standard old/new recognition test. In other words, unlike in standard recognition tests where participants see a word and have to judge whether the word is old or new, in the forced recognition task participants see two words and have to determine which one was presented in the training phase. There were six pairings in the test phase: HO with HN, HO with LN, HO with LO, HN with LO, HN with LN and LO with LN (L, H, O, and N stand for High Frequency, Low Frequency, Old and New respectively). There were four conditions where one old word and one new word were presented, one condition in which two old words were presented, and one condition in which two new words were presented. This method was used because it allowed the researchers to directly compare words between conditions. For example, by giving the test items individually, participants might correctly identify two new words as being new. However, by forcing the participant to choose one of the two words, they are forced to pick the one that feels more familiar. This allowed the researchers to directly compare the HN condition to the LN condition.

The researchers argued that certain conditions had to be met in order to observe a mirror effect. First, the LO with LN pairing should yield the highest rate of correct responses relative to the other pairings with an old and a new item. This is because, as discussed above, the $d'$ (the discriminability between the old and the new words) is
largest between the low frequency old words and the low frequency new words making words from these two conditions the easiest to discriminate. Second, the HO-HN pairing should yield the lowest rate of correct responding of the old plus new pairings. This is because in this condition, the $d'$ is smallest making the words from these two conditions more difficult to discriminate. Third, the accuracy of the LO-HN and the HO-LN pairings should lie somewhere in between the accuracy of the LO-LN and HO-HN pairings.

Finally, the researchers argued that the probability of choosing the low frequency old word over the high frequency old word should be more than 0.5, and the probability of choosing a high frequency new word over a low frequency new word should also be more than 0.5.

The principal manipulation made by Glanzer, Adams, and Iverson (1991) was the delay between the training phase and the test phase. More specifically, there was an immediate and a delayed condition. This manipulation was made both within subject (in experiment 1) and between subject (in experiment 2). They argued that by delaying the time between the training phase and the test phase, the number of features that became marked in the experimental session would be reduced. Intuitively, one would expect this type of manipulation to have an effect on only the words studied in the training phase because a delay between the training phase and the test phase should not influence the familiarity of the new items. It would be possible that with a lower threshold more new items would be called old, but the difference between the high frequency new words and the low frequency new words should remain constant. In other words, an intuitive prediction would be that the difference between the high frequency old words and the low frequency old words might be reduced, but the difference between the new words should
not change. This is not the prediction, however, that Glanzer, Adams, and Iverson made. They argued that as time passes, features of the words that were marked in the training phase of the experiment become unmarked. In other words as the delay between the training and the test increases it becomes more difficult to tell if the words were encountered in the training phase or not. The same proportion of marked features will become unmarked for the high and the low frequency words. Thus, as time passes, the hits for the low frequency words will remain superior to the hits for the high frequency words, but the difference between the two will decline. However, because the probability of correctly identifying a new word is tied to the probability of correctly identifying an old word, the difference between the high and low new words will also be reduced. Thus, with a delay between the training and the test, the false alarms for the high frequency words will remain higher to the false alarms for the low frequency words, but the difference between the two will be reduced.

The results for both the between and within-subject designs confirmed Glanzer, Adams, and Iverson’s (1991) prediction. The mirror effect was observed in both the immediate and the delayed conditions. The main prediction that was made was that the delay should lead to the difference between the high and the low frequency new words being reduced. In the condition in which two new words were presented (one high frequency and one low frequency), the proportion of trials in which participants chose the low frequency word was .65 in the immediate condition compared to .57 in the delayed condition. Glanzer, Adams and Iverson concluded that because the old distributions are tied to the new distributions, they must be based on a single process.
The attention likelihood model is illustrated in Figure 1. According to this model, any experimental manipulation that decreases accurate remembering, such as delay between the training phase and the test phase, will also have an effect on the distributions of the new words.
CHAPTER 2: TESTING THE MODELS

As stated previously, the goal of this thesis is not to determine which of the models are correct and which are incorrect. The primary goal of this research is to further our understanding of the word frequency mirror effect. While I do not expect to prove or to disprove any one account, I will critically evaluate how well each of the models accounts for the data found.

There are two primary questions that I would like to investigate. The first is whether a single process account or a dual process account best explains the word frequency mirror effect. As discussed above, there is evidence for both perspectives. In one case, dissociations between hits and false alarms are seen as evidence in favour of a dual process model; in the other case, a lack of dissociation between hits and false alarms is seen as evidence in favour of a single process model. The experiments presented in this thesis will attempt to answer this question by determining under what conditions the hit portion of the mirror effect can be dissociated from the false alarm portion of the mirror effect.

The second question that will be investigated is: what is the difference in the participant’s subjective experiences when encountering high and low frequency words during the training phase? In other words, are participants aware of the difference between the high and the low frequency words? If so, do participants adjust the quantity or the quality of their study behaviour depending on the frequency of the stimuli? I am not suggesting that people are consciously aware that they are being presented with high and
low frequency words; I am simply suggesting that the processing experience that the
participants have might be different for the high and the low frequency words. This
processing difference may lead participants to differentially encode the stimuli. If a word
is studied more, then this might lead to more attention and thus more familiarity for the
low frequency words during the test phase according to the single process model. In
terms of a dual process account, these additional study resources may lead to the low
frequency words being more recollectable.

Both questions will be investigated by using metacognitive measures during the
training phase of the experiments. Metacognition refers to thoughts about one’s own
cognitive processes. For example, metamemory refers to people’s thoughts about their
own memories. A common metamemory procedure is the Judgment of Learning (JOL)
procedure. In this procedure a series of word pairs are presented in the training phase of
an experiment. Then, in a JOL phase, participants are given one word of the word pair,
and are then asked to rate the likelihood that the paired word will be remembered on a
later memory test.

The JOL procedure has been widely used and one finding of particular interest is
the delayed JOL effect (Nelson & Dunlosky, 1991). This effect is the observation that
JOLs are far more accurate when there is a delay between the initial learning of a word
pair and the JOL. When JOLs are made immediately, they tend to be inaccurate. This
could have implications for how participants study words during the training phase of a
recognition experiment. If participants allocate study time and attentional resources to
words based on a metacognitive judgment made as they process the item, then it is
probable that the allocation will be based on a prediction that will ultimately prove to be incorrect.

In order to predict how participants will respond to training items in the context of a mirror effect study, it is important to know what people base their JOL decisions on. Current research on JOLs demonstrates that they are inferential and based on heuristics (Begg, Duft, & Lalonde, 1989; Matvey, Dunlosky & Guttentag, 2001; Koriat, Bjork, Shefer, & Bar, 2004). There are two main theories concerning how the heuristics are used in order to make a JOL. The first theory was proposed by Koriat (1997). According to this theory, a JOL can be based on intrinsic, extrinsic and mnemonic cues. An intrinsic cue is a cue that is inherent in the item, for example word frequency. An extrinsic cue is a cue brought about by external factors not inherent in the item, for example the experimental condition in which the word appears. A mnemonic cue is an evaluation by the learner about how well the item is known. The mnemonic cues are affected by things like cue familiarity (Metcalf, Schwartz, & Joaquim, 1993) and retrieval fluency (Benjamin, Bjork & Schwartz, 1998). Koriat argues that the intrinsic and extrinsic cues can directly affect the JOL, or they can modify the mnemonic cues that will affect the JOLs. He argues that the delayed JOL effect occurs because over time there is an increased reliance on the mnemonic cues as a result of the intrinsic and extrinsic cues being no longer accessible.

A second view proposed by Son and Metcalfe (2005) is that the JOLs are a two step process. First, participants evaluate the familiarity of the cue. If familiarity is low, the item will be given a low JOL. If the familiarity is not low, then the retrieval fluency of the cued word is evaluated. According to this approach, the delayed JOL effect occurs
because immediately both the cue familiarity and the retrieval fluency will be very high, but over time, the cue familiarity and retrieval fluency will only be high for items that are likely to be actually remembered. As such, the delayed JOLs will be more accurate than the immediate JOLs. Thus, in Koriat’s (1997) approach the JOL is formed in one single step that is based on a variety of cues, and in Son and Metcalfe’s approach, the JOL is formed in two steps.

In both cases however, there is a strong reliance on the heuristics and inference in making the JOLs. In general, it has been found that the more fluently an item is processed, the higher the JOL will be for that item (Koriat & Ma’ayan, 2005, Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Matvey, Dunlosky & Guttentag, 2001; Koriat, Bjork, Shefer, & Bar, 2004). In the context of a word frequency study, because high frequency words have been encountered more often, it follows that they are probably experienced with greater fluency than low frequency words. Thus, it is likely that the high frequency words will be evaluated as being more memorable than the low frequency words.

The rational for using the JOL procedure is that maybe after participants initially examine words in the training phase of an experiment, they make some sort of metacognitive judgment about the item being presented. It is possible that this metacognitive judgment is the basis for allocation of study time and attentional resources to that particular item. According to this line of reasoning, if participants do have different experiences for the high and the low frequency words, then participant’s will evaluate the memorability of the items differently, and subsequently change the way that they study the items.
Furthermore, Son and Metcalfe (2000) conducted a literature review and found that in general, people are likely to spend more time studying items that are given lower JOLs. Thus, if low frequency words are given lower memory evaluations, it follows that these words should be studied for longer periods of time in a self-paced training phase. In an experimenter-paced training phase, it is possible that the participants use a more effective training strategy (such as imagery or thinking about the items in different ways) when studying the low frequency words. As for the high frequency words, because the participants feel that these items are already memorable, it is unnecessary for them to use that effective training strategy because the item is already memorable. It could be that because the low frequency words are studied for longer (in the case of a self-paced training phase) or better (in the case of an experimenter-paced training phase), the low frequency words end up being more recollectable to the participants.

The metacognitive judgments I will be examining are very different from the metacognitive judgments described above. For example, in the delayed JOL procedure, the word pairs are presented on a screen, and then several minutes later the participants are asked to make their metacognitive judgment. It is possible that in memory experiments, the participants are making their metacognitive judgments as they study the words. If this is the case, then heuristics such as retrieval fluency and cue familiarity would not apply because the to-be-remembered word is directly in front of them. Because it is these on-the-fly metacognitive judgments that I want to examine (and these metacognitive judgments are clearly based on different processes than the ones proposed by Koriat (1997) and Son and Metcalfe (2005)), I will not use the term JOL for these judgments. Instead I will use the term Estimations of Remembering (EORs).
As described, there is a large body of evidence that suggests that JOLs are based on heuristics and inference, and it is unlikely that EORs are any different. However, it is the specific heuristics that are used that may differ between the JOLs and the EORs.

The second measure I will use to study the participant's experience during the training phase of a word frequency experiment is an allocation of study time measure. There have been numerous studies conducted on the relationship between JOLs and the allocation of study time (for a review see Son and Metcalfe, 2000). In general it is found that items that are judged as being more difficult are studied for longer amounts of time. This measure will be used for two reasons. First, to determine if people do spend more time studying the low frequency words, and second, to determine if there is a relationship between the amount of study time and accurate recognition.

While it is obvious how these metacognitive measures will help answer the question regarding how participants experience the high and the low frequency words, it is less obvious how using these measures will determine if the hits can be dissociated from the false alarms.

An interesting interpretation of the delayed JOL effect was provided by Spellman and Bjork (1992). They argued that the act of giving an item a high JOL actually increased the probability that an item would be recalled in the final test phase. In other words, they argued that the metacognitive judgment affected the recollectability of the items. Evidence for this idea was provided by Kimball and Metcalfe (2003). While the processes that underlie an EOR may differ from the processes that underlie the JOLs, it is possible that the EORs could have the same effect as the JOLs.
To determine if the EORs will affect the recollectability of the items, the processes that underlie the EORs must first be speculated. Unlike the JOLs, an availability heuristic cannot be used because there is no delay between the studying of the word and the EOR. Furthermore, a cue familiarity heuristic cannot be used because there is no cue presented when performing the EOR. One possible cue that the participants could use when performing an EOR is a conceptual fluency heuristic. In other words, participants might think about how much they know about the words when they are performing their EORs. If this is the case, then the EORs might increase the probability of recognition because the participants had studied the items in a more elaborate way. Thus the hit portion of the mirror effect would be affected. If the hit portion of the mirror effect is affected, then a test between the single and dual process accounts of the mirror effect can be performed. The single process model argues that the hits and the false alarms are related and thus if there is any change in one, there should be a change in the other. However, the dual process model argues that the hits and the false alarms are dissociated and thus a change in one does not necessarily have to be accompanied by a change in the other.

**Experiment 1: EORs and Self Paced Training**

The first experiment was designed to answer the two research questions posed earlier. The first goal was to determine if the hit portion of the mirror effect can be dissociated from the false alarm portion. The second to determine if there was a relationship between the EORs and the study times of the words. To answer these questions, a standard recognition memory experiment was conducted in which high and low frequency words were presented in the training phase of the experiment. The study
phase was self-paced such that each participant could study each word for as long as they felt they needed in order to learn the word. After they had studied the word, the participants had to make an EOR judgment on a scale from one to six. The data was analyzed looking at the overall hit and false alarm rates, as well as the overall study time data. A median split analysis was conducted separating the items studied for a long period of time from the items studied for a short period of time. The median split analysis was conducted in such a way that the high frequency words were separated from the low frequency words, then the study times were rank ordered within each condition. The fast half of study times were then separated from the slow half of study times in each condition. This median split analysis was performed for every participant. Thus, there were four experimental groups with equal numbers of trials in each: high frequency words studied for a short period of time, high frequency words studied for a long period of time, low frequency words studied for a short period of time, and low frequency words studied for a long period of time. By looking at the median split data, it can be determined if there is a difference between the words that are studied for longer periods of time versus words that are studied for shorter periods of time in terms of the hit rates and the EOR judgments.

A similar median split analysis was conducted separating the items rated as being memorable from the items rated as not memorable. This median split was performed in the same manner as the study time median split described above. By looking at the EOR median split data, it can be determined if there is a difference between the words that are rated as being memorable from the items rated as not memorable in terms of the hit rates and the study times.
Method

Participants. Forty undergraduates from Simon Fraser University participated in the experiment for course credit or for an entry into a lottery. All participants had normal or corrected-to-normal vision.

Stimuli. Sixty high frequency words and sixty low frequency nouns were used as the training and the test stimuli. The stimuli were chosen from the Kucera-Francis (1967) database of word frequency. The low frequency words had a frequency of between 15 and 50 uses/million and the high frequency words had a frequency of between 100 and 150 uses/million. There was no significant difference between the word length or the number of syllables between the high and the low frequency words.

Procedure. During the training phase, participants saw thirty high frequency and thirty low frequency words in the middle of a computer screen chosen at random from the original list of 120. Each item was presented a maximum of once during the training phase. The words appeared on the screen one at a time in random order. The items appeared in the middle of the computer screen (upper case, black, 24 point, Monaco font). The participants were instructed to study each word for as long as they needed in order to learn the word for a later memory test. Furthermore, the participants were instructed to use their time in the most efficient way possible, in other words to spend more time studying words they thought would be hardest to remember. After studying each word, the participants pressed a button to continue. Once the button had been pressed, the word disappeared from the screen and the question “On a scale from 1 to 6, how likely is it that you will remember this word on a later memory test?” appeared in the middle on the screen. Participants entered their response on a keyboard. After they had done so, the next trial began. The EORs and study times were recorded for analysis.
During the test phase, all 120 words were presented one at a time in the middle of a computer screen in random order. When each word was presented, the phrase “Did you see this word in the Training Phase? Yes or No?” was presented below the item. To respond “Yes” participants pressed one button on a response box, to respond “No”, they pressed another button on the same response box. The hit and false alarm data were recorded.

**Results**

The data from two participants were excluded from analysis because their overall accuracy was less than 75%. It should be noted that due to a programming error, the EORs were not properly recorded for 4 participants. Thus, the overall accuracy data is based on all 38 participants, but the overall EOR data and the median split data is only based on 34 participants.

Alpha in these experiments was set at 0.05. P-values under this value were considered significant, whereas p-values over this value were considered non significant. However, not all significant effects were treated in the same manner. In the six experiments hundreds of F and T tests were performed and as such, Type 1 errors need to be considered. Because many of the experiments had similar comparisons, an effect that was replicated across experiments is worth discussing more than an effect that is only observed in one experiment.

**Overall Data.** Overall results are presented in Table 1. A 2 X 2 ANOVA was performed to look at the hit and false alarm data. The variables of the ANOVA were: frequency (high vs. low) and claiming oldness (hits vs. false alarms). Collapsed across frequency, participants achieved 90% hits and 9% false alarms, $F(1,37) = 2159.16$, $MSE$
= .012, \( p < .001 \). Collapsed across hits and false alarm, participants were more likely to claim the high frequency words as being old, \( F(1,37) = 16.226, MSE = .003, p < .001 \).

There was a frequency by oldness interaction observed, \( F(1,37) = 19.276, MSE = .003, p < .001 \). Examining the interaction further, it was found that high frequency new words were more likely to be falsely recognized, \( t(37) = 6.04, p < .001 \). However, there was no difference in correct recognition between the high and low frequency words, \( t(37) = .000, p = 1.000 \).

Overall, high frequency words were rated as being more memorable than low frequency words, \( t(33) = 5.16, p < .001 \). However, there was no difference between the high and the low frequency words in terms of how long the words were studied, \( t(37) = .063, p = .949 \).
Table 1

Experiment 1: P(old), EORs, and Study Times for High and Low Frequency Words

<table>
<thead>
<tr>
<th>Frequency</th>
<th>p(old)</th>
<th>Study time (ms)</th>
<th>EOR</th>
<th>p(old)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>.90 (.01)</td>
<td>5843 (803)</td>
<td>3.85 (.13)</td>
<td>.12 (.01)</td>
</tr>
<tr>
<td>Low</td>
<td>.90 (.01)</td>
<td>5831 (714)</td>
<td>3.48 (.11)</td>
<td>.05 (.01)</td>
</tr>
</tbody>
</table>

Note. p(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

Study Time Median Split Data. Study time median split data are presented in Table 2. To examine the relationship between study time, EORs and frequency, two 2 X 2 ANOVAs were conducted; one was conducted on the hit data and one was conducted on the EOR data. The variables of all three ANOVAs were: frequency (high vs. low) and study time (short vs. long). Non-significant interactions and comparisons (p > .1) will not be reported.

First, looking at the hit data, as described above, the low frequency words were not more likely to be correctly recognized than the high frequency words. There was no reliable effect of study time on the hits, F(1,33) = .371, MSE = .009, p = .371, meaning that the items that were studied for longer were not recognized more accurately. There was a reliable frequency by study time interaction, F(1,33) = 4.473, MSE = .006, p = .042. Examination of the interaction revealed that for the low frequency words, the items
studied for longer were remembered better, \( t(33) = 2.05, p = .049 \), but this relationship did not hold up for the high frequency words, \( t(33) = .653, p = .518 \).

Next, looking at the EOR data, there was no difference in the EORs between items studied for a long period of time compared to items studied for a short period of time, \( F(1,33) = .101, MSE = .406, p = .752 \).

Table 2
Experiment 1: \( p(\text{old}) \), EORs and Study Times for High and Low Frequency Words
Median Split by Study Time

<table>
<thead>
<tr>
<th>Frequency</th>
<th>( p(\text{old}) )</th>
<th>Study time (ms)</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.90 (.02)</td>
<td>2975 (411)</td>
<td>3.91 (.17)</td>
</tr>
<tr>
<td>Long study time</td>
<td>.89 (.02)</td>
<td>8971 (1389)</td>
<td>3.78 (.13)</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.87 (.02)</td>
<td>3086 (368)</td>
<td>3.44 (.14)</td>
</tr>
<tr>
<td>Long study time</td>
<td>.91 (.02)</td>
<td>8857 (1223)</td>
<td>3.51 (.12)</td>
</tr>
</tbody>
</table>

Note. \( p(\text{old}) = \) proportion of items claimed as "old". Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

EOR Median Split Data. EOR median split data are presented in Table 3. To examine the relationship between study times, EORs and frequency, two 2 X 2 ANOVAs were conducted; one was conducted on the hit data and one was conducted on the study
time data. The variables of all three ANOVAs were: frequency (high vs. low) and EOR (low vs. high). Non-significant interactions and comparisons ($p > .1$) will not be reported.

First, looking at the hit data, as described above, the low frequency words were not more likely to be correctly recognized than the high frequency words. There was a reliable effect of EORs on accuracy, $F(1,33) = 33.75, MSE = .008, p < .001$, indicating that the items that were rated as being more memorable, were more likely to be correctly recognized. Examination of the EOR effect revealed that for both the high and the low frequency words, the items given higher EORs were more accurately remembered, $t(33) = 4.46, p < .001$, and $t(33) = 4.28, p < .001$ respectively.

In terms of study times, items given higher EORs ended up being studied for longer, $F(1,33) = 10.90, MSE = 4314472.18, p = .002$. The finding that words given high EOR were studied for longer than words given low EORs was consistent for both high frequency words, $t(33) = 2.11, p = .042$, and for low frequency words, $t(33) = 4.20, p < .001$. 


Table 3  
*Experiment 1: P(old), EORs and Study Times for High and Low Frequency Words*  
*Median Split by EOR*

<table>
<thead>
<tr>
<th>Frequency</th>
<th>p(old)</th>
<th>Study time (ms)</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low EOR</td>
<td>.85 (.02)</td>
<td>5543 (800)</td>
<td>2.93 (.14)</td>
</tr>
<tr>
<td>High EOR</td>
<td>.94 (.01)</td>
<td>6503 (1015)</td>
<td>4.77 (.15)</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low EOR</td>
<td>.85 (.02)</td>
<td>5326 (747)</td>
<td>2.43 (.11)</td>
</tr>
<tr>
<td>High EOR</td>
<td>.94 (.01)</td>
<td>6618 (837)</td>
<td>4.53 (.14)</td>
</tr>
</tbody>
</table>

Note.  p(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

**Discussion**

Several interesting findings are observed in these data. First, the hit portion of the word frequency mirror effect was not observed, while the false alarm portion was observed. This is interesting because it suggests that the hit portion of the effect may be independent from the false alarm portion of the effect. This finding is consistent with dual process accounts of recognition memory. These dual process accounts would argue that the recollection advantage of the low frequency words was overcome by causing the high frequency words to become more recollectable. While the exact means by which the high frequency words became more recollectable may not be known, what is important is the dissociation between the hits and the false alarms.
Another noteworthy finding is the differences between the EORs and the hits. Items that were rated as being more memorable ended up being more likely to elicit a hit. What is not known from this data is the direction of this relationship. One possibility is that participants do in fact know what words will be remembered in the test phase of the experiment. Another possibility is that rating an item as being memorable actually causes that item to become more memorable. This causal relationship could occur in two ways. First, rating an item as being memorable may cause participants to study the item more in order for their predictions to be correct. This might occur because participants think about their EOR as they study the item. Another possibility is that thinking about the cues on which the EOR is based may increase memory for the item. In other words, an item may be given a high EOR if they can think about many of the conceptual properties of the item. However, thinking about the conceptual properties of the item might actually cause the item to become more memorable. Thus, items that participants are able to think more about will be given higher EORs and will become more recollectable.

Finally, items that were studied for a longer period of time were not more likely to be correctly recalled. This seems to suggest that the quality of learning, not the quantity, is what is important for later remembering.

**Experiment 2: Self-Paced Training**

Although the findings from Experiment 1 deviate from the standard word frequency mirror effect, calling these findings a dissociation might be a misnomer because there is no control condition using the exact same procedure and stimuli. The first reason Experiment 2 was conducted was to demonstrate a standard mirror effect using the same stimuli and similar procedure.
Furthermore, while the hit portion of the word frequency mirror effect was eliminated in Experiment 1, the precise cause of that finding is unclear. The training phase of Experiment 1 differed from a standard word frequency mirror effect in two respects: the self-paced training and the EOR. Experiment 2 was conducted to determine if the EOR was critical to the elimination of the hit portion of the word frequency mirror effect. Experiment 2 was identical to Experiment 1, except that participants did not perform an EOR after studying the items.

**Method**

*Participants.* Twenty-four undergraduates from Simon Fraser University participated in the experiment for course credit or for an entry into a lottery. All participants had normal or corrected-to-normal vision.

*Stimuli.* The same materials were used in Experiment 2 as were used in Experiment 1.

*Procedure.* During the training phase, participants saw thirty high frequency and thirty low frequency words in the middle of a computer screen chosen at random from the original list of 120. Each item was presented a maximum of once during the training phase. The words appeared on the screen one at a time in random order. The items appeared in the middle of the computer screen (upper case, black, 24 point, Monaco font). The participants were instructed to study each word for as long as they needed in order to learn the word for a later memory test. Furthermore, the participants were instructed to use their time in the most efficient way possible, in other words to spend more time studying words they thought would be hardest to remember. After studying
each word, the participants pressed a button to continue. Once the button had been pressed, the word disappeared from the screen. The study time was recorded for analysis.

The test phase was identical to Experiment 1.

Results

**Overall Data.** Overall results are presented in Table 4. The data from two participants was excluded from analysis because their overall accuracy was less than 75%. A 2 X 2 ANOVA was performed to look at the hit and false alarm data. The variables of the ANOVA were: frequency (high vs. low) and oldness (hits' vs. false alarms). Collapsed across frequency, participants achieved 86% hits and 9% false alarms, $F(1,21) = 583.63, MSE = .022, p < .001$. Collapsed across hits and false alarms, there was no significant difference between the high and the low frequency words in terms of the probability of claiming to have seen the words in the training phase, $F(1,21) = 3.231, MSE = .022, p = .087$. There was a significant interaction between oldness and frequency, $F(1,21) = 26.161, MSE = .002, p < .001$. Examining the interaction further, the low frequency words were significantly more likely to elicit a hit, $t(21) = 4.45, p < .001$, and the high frequency words were more likely to elicit a false alarm, $t(21) = 2.35, p = .029$. Thus, a standard mirror effect was observed.

Turning to the study time data, the low frequency words were studied for longer than the high frequency words, $t(21) = 2.63, p = .016$.  

47
### Table 4
**Experiment 2: *P*(old), EORs, and Study Times for High and Low Frequency Words**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>p</em>(old)</td>
<td>Study time (ms)</td>
</tr>
<tr>
<td>High</td>
<td>.82 (.02)</td>
<td>9109 (1086)</td>
</tr>
<tr>
<td>Low</td>
<td>.90 (.02)</td>
<td>9900 (1192)</td>
</tr>
</tbody>
</table>

**Note.** *p*(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. Standard errors are reported in parentheses.

**Median Split Data.** Median Split data are presented in Table 5. To examine the relationship between study time and frequency, a 2 X 2 ANOVAs were conducted; one was conducted on the hit data, one was conducted on the study time data. The variables the ANOVA was: frequency (high vs. low) and study time (short vs. long).

First, looking at the hit data, there was no significant effect of study time on the hit data, *F*(1,21) = .542, *MSE* = .016, *p* = .470, meaning that the items that were studied for longer were not recognized more accurately than items that were studied for a short period of time. There was no reliable frequency by study time interaction, *F*(1,21) = 4.068, *MSE* = .004, *p* = .057. Examination of the data revealed that low frequency words were better recognized than high frequency words both for words that were studied for a short period of time *t*(21) = 4.54, *p* < .001, and for words that were studied for a long period of time, *t*(21) = 2.14, *p* = .044.
Table 5

Experiment 2: P(old) and Study Times for High and Low Frequency Words Median Split by Study Time

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Short study time</th>
<th>Long study time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p(old)</td>
<td>Study time (ms)</td>
</tr>
<tr>
<td>High</td>
<td>.80 (.03)</td>
<td>4385 (591)</td>
</tr>
<tr>
<td>Low</td>
<td>.90 (.02)</td>
<td>4951 (636)</td>
</tr>
</tbody>
</table>

Note. p(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. Standard errors are reported in parentheses.

Discussion

In Experiment 2, a standard mirror effect was observed. This is interesting as it suggests that the hit portion of the word frequency mirror effect was eliminated in Experiment 1 because of the inclusion of the EOR. Yet another possibility is that the hit portion of the mirror effect was eliminated in Experiment 1 because of an interaction between the EOR and the self paced study time. This possibility was examined in an Experiment not presented in this thesis (but still worth mentioning) by having participants perform a EOR after a fixed study time. In this case, the results were identical to the results of Experiment 1. Taken together, these experiments suggest that inclusion of the metacognitive judgment eliminates the hit portion of the mirror effect.

Furthermore, Experiment 2 provides a control condition for Experiment 1. Taken together, it is observed that the hit portion of the word frequency mirror effect can be manipulated independently from the false alarms. Again this seems to support a dual process account of the word frequency mirror effect although, as will be discussed in the
General Discussion, single process accounts might be able to account for this dissociation with some minor adjustments.

Consistent with the findings of Experiment 1, the items that were studied for longer were not more likely to elicit a hit. However, inconsistent with the findings of Experiment 1 (and with all other to be presented experiments) the low frequency words were studied for longer than the high frequency words. This finding is difficult to interpret as this is the only experiment in which the high and low frequency words differ in terms of study time (i.e., it is not a reliable finding). Because there were hundreds of F and T tests conducted, it is difficult to interpret any effect that is not replicated across experiments.

Experiment 3: Pleasantness Ratings

It appears that the EORs are affecting the hit portion of the mirror effect. Not determined by the first two experiments is how the EORs are affecting the mirror effect. The goal of this experiment was to determine if making an evaluation about memory is the critical component of the EOR that eliminates the hit portion of the mirror effect, or if it is any type of rating that will eliminate the mirror effect. To answer this question, an experiment identical to Experiment 1 was conducted with one modification. Instead of rating the memorability of the words, participants had to rate how pleasant they found each word. Thus, if the hit advantage for the low frequency words returns, it would indicate that evaluating one's memory is what is important in eliminating the hit portion of the mirror effect. If however, the hit advantage for the low frequency words is not observed, it would indicate that evaluating one's memory is not important in eliminating
the hit portion of the mirror effect; instead it is making any type of rating that eliminates the effect.

In addition to answering that question, the same questions involving the pleasantness ratings and the study times can be examined. The data were analyzed looking at the overall hit and false alarm rates as well as the overall pleasantness ratings and study times. Furthermore, the relationship between the study time and the hits, the pleasantness ratings and the hits, and the study time and pleasantness ratings can all be examined by using the median split analysis procedure described in Experiment 1.

**Method**

*Participants.* Twenty-two undergraduates from Simon Fraser University participated in the experiment for course credit or for an entry into a lottery. All participants had normal or corrected-to-normal vision.

*Stimuli.* The same materials were used in Experiment 3 as were used in Experiment 1.

*Procedure.* During the training phase, participants saw thirty high frequency and thirty low frequency words in the middle of a computer screen chosen at random from the original list of 120. Each item was presented a maximum of once during the training phase. The words appeared on the screen one at a time in random order. The items appeared in the middle of the computer screen (upper case, black, 24 point, Monaco font). The participants were instructed to study each word for as long as they needed in order to learn the word for a later memory test. Furthermore, the participants were instructed to use their time in the most efficient way possible, in other words to spend more time studying words they thought would be hardest to remember. After studying
each word, the participants pressed a button to continue. Once the button had been pressed, the word disappeared from the screen and the question “On a scale from 1 to 6, how pleasant do you find this word?” appeared in the middle on the screen. Participants entered their response on a keyboard. After they had done so, the next trial began. The pleasantness ratings and study times were recorded for analysis.

The test phase was identical to Experiment 1.

Results

Overall Data. Overall results are presented in Table 6. A 2 X 2 ANOVA was performed to look at the hit and false alarm data. The variables of the ANOVA were: frequency (high vs. low) and oldness (hits vs. false alarms). Collapsed across frequency, participants achieved 90% hits and 8% false alarms, $F(1,21) = 1562.12$, $MSE = .010$, $p < .001$. Furthermore, collapsed across hits and false alarms, there was no difference between the probability of claiming a high frequency words as being old and the probability of claiming a low frequency word as being old, $F(1,21) = .823$, $MSE = .003$, $p = .374$. There was a frequency by oldness interaction observed, $F(1,21) = 10.113$, $MSE = .005$, $p = .005$. Examining the interaction further, it was found that high frequency new words were more likely to be falsely recognized, $t(21) = 2.63$, $p = .016$. However, the low frequency words were more likely to be correctly recognized, $t(21) = 2.26$, $p = .035$. That is, standard mirror effect was observed.

Overall, high frequency words were rated as being more pleasant than low frequency words, although the result was not significant, $t(21) = 1.85$, $p = .079$. Also, the low frequency words were studied for longer, but this result was also not significant, $t(21) = 1.93$, $p = .067$. 

52
Table 6
Experiment 3: *p*(old), Pleasantness Ratings (PR), and Study Times for High and Low Frequency Words

<table>
<thead>
<tr>
<th>Frequency</th>
<th><em>p</em>(old)</th>
<th>Study time (ms)</th>
<th>PR</th>
<th><em>p</em>(old)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>.89 (.02)</td>
<td>3949 (640)</td>
<td>3.92 (.17)</td>
<td>.11 (.02)</td>
</tr>
<tr>
<td>Low</td>
<td>.92 (.02)</td>
<td>4646 (711)</td>
<td>3.78 (.16)</td>
<td>.05 (.01)</td>
</tr>
</tbody>
</table>

Note. *p*(old) = proportion of items claimed as “old”. PR = Pleasantness Rating (on a scale from 1 to 6). Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. Standard errors are reported in parentheses.

*Study Time Median Split Data.* Study time median split data are presented in Table 7. To examine the relationship between study time, EORs and frequency, two 2 X 2 ANOVAs were conducted; one was conducted on the hit data and one was conducted on the pleasantness data. The variables of both ANOVAs were: frequency (high vs. low) and study time (short vs. long). Most non-significant interactions and comparisons (*p* > .1) will not be reported.

There was no reliable effect of study time on the hits, *F*(1,21) = .000, *MSE* = .006, *p* = 1.000, meaning that the items that were studied for longer were not recognized more accurately.

Looking at the pleasantness data, there was no difference in the pleasantness rating for items studied for a long time compared to items studied for a short period of time, *F*(1,21) = .232, *MSE* = .267, *p* = .635.
Table 7

Experiment 3: *p*(old), Pleasantness Ratings (PR) and Study Times for High and Low Frequency Words Median Split by Study Time

<table>
<thead>
<tr>
<th>Frequency</th>
<th>p(old)</th>
<th>Study time (ms)</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.88 (.03)</td>
<td>2191 (387)</td>
<td>3.97 (.21)</td>
</tr>
<tr>
<td>Long study time</td>
<td>.89 (.02)</td>
<td>5708 (920)</td>
<td>3.87 (.15)</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.93 (.02)</td>
<td>2416 (471)</td>
<td>3.78 (.19)</td>
</tr>
<tr>
<td>Long study time</td>
<td>.92 (.02)</td>
<td>6876 (1003)</td>
<td>3.77 (.17)</td>
</tr>
</tbody>
</table>

Note. *p*(old) = proportion of items claimed as “old”. PR = Pleasantness Rating(on a scale from 1 to 6). Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. Standard errors are reported in parentheses.

Pleasantness Median Split Data. Pleasantness median split data are presented in Table 8. To examine the relationship between study time, pleasantness ratings and frequency, a 2 X 2 ANOVAs were conducted on the hit data. The variables the ANOVA was: frequency (high vs. low) and pleasantness ratings (low vs. high). Non-significant interactions and comparisons (*p*. 1) will not be reported.

As described above, the low frequency words were more likely to be correctly recognized than the high frequency words. There was a reliable effect of pleasantness ratings on the hits, $F(1,21) = 10.25$, $MSE = .005$, $p = .004$, meaning that the items that were rated as being more pleasant, were remembered more accurately. Examination of the pleasantness effect on the hits revealed that for the high frequency words, the items
given higher pleasantness ratings were more likely to be accurately recognized than items given low pleasantness ratings, \( t(21) = 2.47, p = .022 \). The difference in the hits between the low frequency words given high and low pleasantness ratings was in the same direction as the high frequency words, but was not significant, \( t(21) = 1.86, p = .077 \).

Also, there were significantly more hits on the low frequency words for items given low pleasantness ratings, \( t(21) = 2.09, p = .050 \), but there was no significant difference between high and low frequency items for items given high pleasantness ratings, \( t(21) = 1.02, p = .318 \).

Table 8

<table>
<thead>
<tr>
<th>Frequency</th>
<th>p(old)</th>
<th>Study time (ms)</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low PR</td>
<td>.85 (.03)</td>
<td>4038 (683)</td>
<td>2.86 (.17)</td>
</tr>
<tr>
<td>High PR</td>
<td>.92 (.03)</td>
<td>3861 (642)</td>
<td>4.98 (.19)</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low PR</td>
<td>.91 (.02)</td>
<td>4606 (719)</td>
<td>2.65 (.17)</td>
</tr>
<tr>
<td>High PR</td>
<td>.94 (.02)</td>
<td>4686 (746)</td>
<td>4.91 (.18)</td>
</tr>
</tbody>
</table>

Note. p(old) = proportion of items claimed as “old”. PR = Pleasantness Rating (on a scale from 1 to 6). Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. Standard errors are reported in parentheses.

Discussion

First looking at the hit data, the low frequency advantage on the hit portion of the word frequency mirror effect was observed. This would seem to indicate that in
Experiment 1, it was the evaluations of the subject's memory in the training phase of the experiment that was important for eliminating the hit portion of the mirror effect. It would appear that way because in this experiment, when participants are asked to provide a non-memory related rating, the low frequency advantage on the hit portion of the mirror effect was present.

Overall there was no reliable difference in the pleasantness ratings between the high and the low frequency words, although the difference was close to significant. Overall it was the high frequency words that were rated as being more pleasant. Although this is a different rating, this finding follows the trend of Experiment 1. As will be discussed in the General Discussion, it is possible that participants are using the same heuristic for both ratings.

There was no reliable difference in the study times between the high and low frequency words, although the difference approached significance. This data follows the trend of Experiment 2 in which the low frequency words were studied for longer than the high frequency words. It might be tempting to conclude that the low frequency hit advantage arises from the longer study times; however, just as in Experiment 2, there was no difference in the hits between the words studied for a long period of time and the words studied for a short period of time. Another consistency to keep in mind is that in Experiment 2 and the present experiment, the low frequency words were studied for longer than the high frequency words, and a mirror effect was observed. In Experiment 1, there was no difference in the study times between high and low frequency words and in this experiment there was no low frequency advantage in the hits.
**Experiment 4: EOR First**

In the previous experiments, there have been several interesting results. One result has been the appearing and disappearing low frequency advantage on the hit portion of the word frequency mirror effect. Thus far, it has been demonstrated that when participants perform an EOR that is based on their own evaluation of their memory, there is no low frequency advantage on the hit portion of the mirror effect. Experiment 4 and 5 were conducted to explore why the EORs had this effect.

One possible reason is that the cues and heuristics used by participants to perform their EORs also affect the memory of these items. If participants use a conceptual fluency heuristic (in other word if they think about how much they know about the item) the high frequency words will elicit higher EORs. However, if this is the case, then the high frequency words will be thought about in several different contexts, and this could lead to an increased likelihood that the high frequency words will elicit a hit thereby eliminating the low frequency word advantage. This hypothesis is supported by the observation that items that are given higher EORs are more likely to elicit a hit.

Experiment 4 was designed to explore this hypothesis. Thinking about the concept of an item requires processing time. In this experiment, participants saw each word for one second before having to make their EOR. After they made their EOR, the word reappeared on the screen and participants could study the word for as long as they wanted. Because the exposure to the item is quite brief, it is possible that the participants will use a different heuristic in order to perform the EOR. If this is the case, then both the EOR pattern and the hit pattern might change from the previous experiments. If this is not the case, then the EOR and hit pattern might remain constant.
Method

Participants. Twenty undergraduates from Simon Fraser University participated in the experiment for course credit or for an entry into a lottery. All participants had normal or corrected-to-normal vision.

Stimuli. The same materials were used in Experiment 4 as were used in Experiment 1.

Procedure. During the training phase, participants saw thirty high frequency and thirty low frequency words in the middle of a computer screen chosen at random from the original list of 120. Each item was presented a maximum of once during the training phase. The words appeared on the screen one at a time in random order. The items appeared in the middle of the computer screen (upper case, black, 24 point, Monaco font). Each word was presented on the computer screen for one second. After one second, the word disappeared and the question “On a scale from 1 to 6, how likely is it that you will remember this word on a later memory test?” appeared. Participants entered their response on the keyboard. After making their response, the word reappeared in the middle of the screen and the participants were instructed to study each word for as long as he or she needed to learn the word for a later memory test. Furthermore, the participants were instructed to use their time in the most efficient way possible, in other words to spend more time studying words they thought would be hardest to remember. After studying each word, the participants pressed a button to continue. Once the button had been pressed, the word disappeared from the screen and the next trial began. The EOR and study time were recorded for analysis.

The test phase was identical to Experiment 1.
Results

Overall Data. The data from one participant was excluded from analysis because their overall accuracy was less than 75%. Overall results are presented in Table 9. A 2 X 2 ANOVA was performed to look at the hit and false alarm data. The variables of the ANOVA were: frequency (high vs. low) and oldness (hits vs. false alarms). Collapsed across frequency, participants achieved 84% hits and 8% false alarms, \( F(1,18) = 596.34, \) \( MSE = .018, p < .001 \). Furthermore, collapsed across hits and false alarms, there was no difference between the probability of claiming a high frequency words as being old and the probability of claiming a low frequency word as being old, \( F(1,18) = 1.017, MSE = .004, p = .327 \). There was a frequency by oldness interaction observed, \( F(1,18) = 15.033, MSE = .006, p = .001 \). Examining the interaction further, it was found that high frequency words were more likely to elicit a false alarm, \( t(18) = 2.68, p = .015 \). It was also found that the low frequency words were more likely to elicit a hit, \( t(18) = 3.31, p = .004 \). Thus, a standard word frequency mirror effect was observed.

Overall, the high frequency words were rated as being more memorable than low frequency words, \( t(18) = 2.28, p = .035 \). However, there was no difference between the high and the low frequency words in terms of how long the words were studied \( t(18) = .034, p = .973 \).
Table 9

Experiment 4: P(old), EORs, and Study Times for High and Low Frequency Words

<table>
<thead>
<tr>
<th>Frequency</th>
<th>p(old)</th>
<th>Study time (ms)</th>
<th>EOR</th>
<th>p(old)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>.80 (.03)</td>
<td>4577 (1396)</td>
<td>3.38 (.13)</td>
<td>.11 (.02)</td>
</tr>
<tr>
<td>Low</td>
<td>.88 (.02)</td>
<td>4581 (1408)</td>
<td>3.10 (.13)</td>
<td>.05 (.01)</td>
</tr>
</tbody>
</table>

Note. p(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

Study Time Median Split Data. Study time median split data are presented in Table 10. To examine the relationship between study time, EORs and frequency, two 2 X 2 ANOVAs were conducted; one was conducted on the hit data and one was conducted on the EOR data. The variables of all three ANOVAs were: frequency (high vs. low) and study time (short vs. long). Non-significant interactions and comparisons (p > .1) will not be reported.

First looking at the hit data, there was no effect of study time, F(1,18) = .146, MSE = .006, p = .707, meaning that the items that were studied for longer were not more likely to be correctly recognized. Low frequency words were more likely to elicit a hit than high frequency words both for items studied for a long period of time, t(18) = 3.22, p = .005, and for the items studied for a short period of time, there was a non-significant difference in the same direction, t(18) = 1.89, p = .076.
Finally, looking at the EOR data, there was no difference in the EORs for items studied for a long time compared to items studied for a short period of time, $F(1,18) = .011, MSE = .255, p = .917$.

Table 10

**Experiment 4: P(old), EORs and Study Times for High and Low Frequency Words Median Split by Study Time**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>$p(old)$</th>
<th>Study time (ms)</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.80 (.04)</td>
<td>3310 (828)</td>
<td>3.38 (.15)</td>
</tr>
<tr>
<td>Long study time</td>
<td>.79 (.03)</td>
<td>7843 (2005)</td>
<td>3.38 (.14)</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.87 (.03)</td>
<td>3356 (897)</td>
<td>3.09 (.14)</td>
</tr>
<tr>
<td>Long study time</td>
<td>.89 (.02)</td>
<td>7806 (1954)</td>
<td>3.12 (.15)</td>
</tr>
</tbody>
</table>

**Note.** $p(old)$ = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

**EOR Median Split Data.** EOR median split data are presented on Table 11. To examine the relationship between study time, EORs and frequency, two 2 X 2 ANOVAs were conducted; one was conducted on the hit data and one was conducted on the study time data. The variables of all three ANOVAs were: frequency (high vs. low) and EOR (low vs. high). Non-significant interactions and comparisons ($p>.1$) will not be reported.
First looking at the hit data, there was a reliable effect of EOR on the hits, $F(1, 18) = 15.66, MSE = .010, p = .001$, meaning that the items that were rated as being more memorable, were actually remembered more accurately.

Next looking at the EOR data, there was a reliable frequency by EOR interaction, $F(1, 18) = 6.49, MSE = .051, p = .020$. Further examination of the data revealed that high frequency items were rated as more memorable than low frequency words for items given low EORs, $t(18) = 2.99, p = .008$. This pattern was consistent for items given high EORs, but the difference was not significant, $t(18) = 1.14, p = .269$.

Finally, looking at the study time data, there was no effect of EORs on study time, $F(1, 18) = 1.554, MSE = 3350480, p = .229$, meaning that there was no difference in the study times for items given high EORs compared to items given low EORs.
Table 11
Experiment 4: P(old), EORs and Study Times for High and Low Frequency Words Median Split by EOR

<table>
<thead>
<tr>
<th>Frequency</th>
<th>p(old)</th>
<th>Study time (ms)</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low EOR</td>
<td>.74 (.04)</td>
<td>5730 (1628)</td>
<td>2.52 (.11)</td>
</tr>
<tr>
<td>High EOR</td>
<td>.85 (.03)</td>
<td>5423 (1182)</td>
<td>4.24 (.18)</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low EOR</td>
<td>.84 (.03)</td>
<td>5951 (1558)</td>
<td>2.11 (.11)</td>
</tr>
<tr>
<td>High EOR</td>
<td>.91 (.02)</td>
<td>4212 (1271)</td>
<td>4.09 (.18)</td>
</tr>
</tbody>
</table>

Note. p(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

Discussion

In this experiment, a standard mirror effect was observed, with the low frequency words being more likely to elicit a hit than the high frequency words. Thus, it seems as though the EORs will only eliminate the hit portion of the mirror effect if the EOR is based on a long processing experience with the items. When the EOR is made after a short exposure to the word, a standard mirror effect is observed. This finding seems to indicate that the EORs that participants make after a brief duration is different from the EORs that participants make after an extended study period. One possibility is that when the EORs follow the self paced study time, participants think about their EOR as they
study the word. In Experiment 4, this may not have been the case because the self paced study period was after the EOR had already been made.

There was no difference in the study time between the high and the low frequency words. This finding is consistent with the previous experiments. Similar to Experiment 1, the high frequency words were rated as being more memorable than the low frequency words. Also items that were given higher EORs ended up eliciting more hits than items given low EORs. Consistent with most of the previous experiments, the items that were studied for longer were not remembered any better than the items studied for a short period of time.

**Experiment 5: Variable Context Training.**

The goal of Experiment 5 was to determine if the amount of conceptual processing had an effect on the mirror effect. To determine this, words were presented three times in one of two conditions. In both cases the to-be-remembered words were presented with a related word. In one condition (the consistent condition), the words were presented with the same word all three times. In the other condition (the variable condition) the to-be-remembered words were presented with a different word on all three occasions. McFarland, Rhodes, and Frey (1979) found that items in the variable condition were remembered better than the items in the consistent condition. They argued that this occurred because when participants have to remember if they saw the items or not, in the variable condition there are three different contexts that can be retrieved. As discussed above, one possible explanation of the elimination of the hit portion of the mirror effect was that when performing an EOR, the high frequency words are thought about in several different contexts, whereas the low frequency words are
thought about in a limited number of contexts. Thus the low frequency words might be in something similar to a consistent condition, and the high frequency words might be in something similar to a variable condition. Following the results of McFarland, Rhodes and Frey, this would increase the elaboration of the high frequency words and would reduce or eliminate the word frequency mirror effect.

**Method**

*Participants.* Twenty-six undergraduates from Simon Fraser University participated in the experiment for course credit or for an entry into a lottery. All participants had normal or corrected-to-normal vision.

*Stimuli.* The same training and test words were used in Experiment 5 as were used in Experiment 1 with one exception. For each word, three semantically related words were generated that were used as contextual words. The semantically related words were chosen from a database in the lab. The database was constructed by having participants generate three associates to the high and the low frequency words. The related items presented in this study were the three most commonly occurring associate for each word.

*Procedure.* During the training phase, participants saw thirty high frequency and thirty low frequency words. The words were chosen at random from the original list of 120. The words were presented on a computer screen one at a time in random order (upper case, black, 24 point, Monaco font). Each word was presented with a context word (lower case, black, 24 point, Monaco font). Participants were told that the word in upper case was the word that should be remembered, and the word in lower case was presented to put the to-be-remembered word in context. All of the words were presented three times throughout the list, so participants saw 180 training pairs. Half of the words were paired
with the same context word three times whereas the other half of the words were paired with a different context word on all three occasions. Each word pair was presented on the screen for three seconds. After three seconds, the word pair was replaced with a new word pair.

The test phase was identical to Experiment 1.

Results

Results are presented in Table 12. The data from one participant was excluded from analysis because their overall accuracy was less than 55%. Overall, participants achieved 77% hits and 6% false alarms, t(24) = 23.97, p < .001. To examine the relationship between frequency and consistency of training, a 2 X 2 ANOVA was conducted. The variables were frequency (high vs. low) and consistency of training (variable vs. consistent). Collapsed across conditions, the low frequency words were more likely to elicit a hit than the high frequency words, F(1, 24) = 9.69, MSE = .014, p = .005. Collapsed across frequency, a consistency of training main effect was observed, F(1, 24) = 5.98, MSE = .016, p = .022 with the items in the variable condition eliciting more hits than the items in the consistent condition. Finally, there was no observed frequency by consistency of training interaction, F(1, 24) = 1.89, MSE = .010, p = .182. Examining the main effects further, it was observed that there was no significant difference between the high frequency variable and the high frequency consistent condition, t(24) = 1.40, p = .173, but the items in the low frequency variable condition were more likely to elicit a hit than the items in the low frequency consistent condition, t(24) = 11.511, MSE = .011, p = .002. Furthermore, there was no significant difference between the high and the low frequency words in the consistent condition, t(24) = 1.02, p
= .316, but in the variable condition, the low frequency words were significantly more likely to elicit a hit than the high frequency words, \( t(24) = 2.87, p = .008 \).

Looking at the new words, it was observed that the high frequency words were incorrectly identified as being old words more often than the low frequency words, \( t(24) = 2.78, p = .010 \).

<table>
<thead>
<tr>
<th>Table 12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 5: P(old) for High and Low Frequency Words with Variable and Consistent Training</strong></td>
</tr>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>High</td>
</tr>
<tr>
<td>Low</td>
</tr>
</tbody>
</table>

Note. \( p(\text{old}) \) = proportion of items called “old”. Standard errors are reported in parentheses.

**Discussion**

In this experiment, similar to the McFarland, Rhodes and Frey (1971) experiment, the items in the variable condition were remembered better than the items in the consistent condition. A mirror effect was observed in both conditions, although the difference between the high and the low frequency words was not significant in the consistent condition. One particularly interesting finding in this experiment was that there was no significant different between the high frequency items in the variable condition and the low frequency items in the consistent condition. While it is difficult to draw strong conclusions from this finding because the conditions vary on two
dimensions, this is the pattern of results that were found in Experiment 1 when no mirror effect was observed. Thus, it is possible that in Experiment 1, the high frequency words are in a variable-like condition, and the low frequency words are in a consistent-like condition.

**Experiment 6: EOR Recognition**

As mentioned above, there have been many T and F tests performed in this thesis, and thus making strong conclusions about a result that has arisen from chance is a concern. Experiment 6 was designed to replicate the findings of Experiment 1. In the first experiment, participants were told to study all of the training items for a later memory test. What is not provided to participants is what type of memory test will be given. It is possible that when told they will be doing a memory test, participants might think that they will be doing a recall test, not a recognition test. In Experiments 1 and 4, participants rated the high frequency words as being more memorable than the low frequency words. Early research looking at word frequency found that although low frequency words are more likely to be correctly recognized, high frequency words are more likely to be recalled (Glanzer & Bowles, 1976). So if participants were doing a recall test, they would be correct in asserting that the high frequency words would be remembered better. Thus, it is possible that the participants are actually consciously aware of what will be more memorable in the later test phase, but they are not correct in claiming that the low frequency words are more memorable because the test is not the expected memory test. It is possible that if participants knew that the test would be a recognition test, the study time and EOR data might be affected.
Experiment 6 had two purposes. First was to determine if participants would respond differently to the training items if they knew the test would be a recognition test and not a recall test. To accomplish this goal, the participants were instructed before seeing the training words that the test phase would be a recognition test and to keep that in mind when they were making their EORs. Furthermore, participants were told what a recognition test was, and what to expect in the training phase. The second goal of Experiment 5, was to replicate the hit pattern obtained in Experiment 1. All the same analyses were performed in this experiment as were performed in Experiment 1.

Method

Participants. Twenty undergraduates from Simon Fraser University participated in the experiment for course credit or for an entry into a lottery. All participants had normal or corrected-to-normal vision.

Stimuli. The same materials were used in Experiment 6 as were used in Experiment 1.

Procedure. During the training phase, participants saw thirty high frequency and thirty low frequency words in the middle of a computer screen chosen at random from the original list of 120. Each item was presented a maximum of once during the training phase. The words appeared on the screen one at a time in random order. The items appeared in the middle of the computer screen (upper case, black, 24 point, Monaco font). The participants were instructed to study each word for as long as they needed in order to learn the word for a later memory test. It was made clear to participants that the test phase of the experiment would involve seeing a word, and trying to remember if that word was presented in the training phase or not. Furthermore, the participants were
instructed to use their time in the most efficient way possible, in other words to spend more time studying words they thought would be hardest to remember. After studying each word, the participants pressed a button to continue. Once the button had been pressed, the word disappeared from the screen and the question “On a scale from 1 to 6, how likely is it that you will remember this word on a later memory test?” appeared in the middle on the screen. Participants entered their response on a keyboard. After they had done so, the next trial began. The EORs and study times were recorded for analysis.

Results

Overall Data. Overall results are presented in Table 13. The data from one participant was excluded from the analyses because their overall accuracy was less than 75%. A 2 X 2 ANOVA was performed to look at the hit and false alarm data. The variables of the ANOVA were: frequency (high vs. low) and oldness (hits vs. false alarms). Collapsed across frequency, participants achieved 86% hits and 9% false alarms, $F(1,18) = 736.03, MSE = .015, p < .001$. Furthermore, collapsed across hits and false alarms, there was no significant difference between the probability of claiming a high frequency words as being old and the probability of claiming a low frequency word as being old, $F(1,18) = 1.943, MSE = .003, p = .180$. There was a frequency by oldness interaction observed, $F(1,18) = 8.34, MSE = .004, p = .010$. Examining the interaction further, it was found that high frequency new words were more likely to be falsely recognized, $t(18) = 3.18, p = .005$. However, there was no significant difference in correct recognition between the high and low frequency old words $t(18) = 1.07, p = .300$. The hit portion of the word frequency mirror effect was not observed in this experiment.
Overall, high frequency words were rated as being more memorable than low frequency words, $t(18) = 3.20, p = .005$. However, there was no difference between the high and the low frequency words in terms of how long the words were studied $t(18) = .212, p = .834$.

### Table 13
**Experiment 6: P(old), EORs, and Study Times for High and Low Frequency Words**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>P(old)</th>
<th>Study time (ms)</th>
<th>EOR</th>
<th>P(old)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>.85 (.03)</td>
<td>5135 (851)</td>
<td>3.75 (.22)</td>
<td>.11 (.03)</td>
</tr>
<tr>
<td>Low</td>
<td>.87 (.02)</td>
<td>5174 (733)</td>
<td>3.48 (.19)</td>
<td>.06 (.01)</td>
</tr>
</tbody>
</table>

Note. P(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

**Study Time Median Split Data.** Study time median split data are presented in Table 14. To examine the relationship between study time, EORs and frequency, two 2 X 2 ANOVAs were conducted; one was conducted on the hit data and one was conducted on the EOR data. The variables of both ANOVAs were: frequency (high vs. low) and study time (short vs. long). Non-significant interactions and comparisons ($p > .1$) will not be reported.

First looking at the hit data, as described above, the low frequency words were not more likely to be correctly recognized than the high frequency words. There was a
reliable effect of study time on the hits, $F(1,18) = 5.04$, $MSE = .005$, $p = .037$, meaning that the items that were studied for longer were more likely to be correctly recognized.

Finally, looking at the EOR data, there was no difference in the EORs for items studied for a long time compared to items studied for a short period of time, $F(1,18) = .035$, $MSE = .136$, $p = .854$.

Table 14

Experiment 6: $p(\text{old})$, EORs and Study Times for High and Low Frequency Words

<table>
<thead>
<tr>
<th>Median Split by Study Time</th>
<th>Frequency</th>
<th>p(\text{old})</th>
<th>Study time (ms)</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.83 (.03)</td>
<td>2811 (497)</td>
<td>3.81 (.20)</td>
<td></td>
</tr>
<tr>
<td>Long study time</td>
<td>.86 (.03)</td>
<td>7458 (1236)</td>
<td>3.78 (.23)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short study time</td>
<td>.85 (.03)</td>
<td>2966 (484)</td>
<td>3.48 (.21)</td>
<td></td>
</tr>
<tr>
<td>Long study time</td>
<td>.89 (.02)</td>
<td>7383 (1011)</td>
<td>3.48 (.20)</td>
<td></td>
</tr>
</tbody>
</table>

Note. $p(\text{old}) = \text{proportion of items claimed as "old". Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.}$

**EOR Median Split Data.** EOR median split data are presented on Table 15. To examine the relationship between study time, EORs and frequency, two $2 \times 2$ ANOVAs were conducted; one was conducted on the hit data and one was conducted on the study
time data. The variables of both ANOVAs were: frequency (high vs. low) and EOR (low vs. high). Non-significant interactions and comparisons ($p>.1$) will not be reported.

First, looking at the hit data, there was a reliable effect of EOR on accuracy, $F(1,18) = 27.74, MSE = .007, p < .001$, meaning that the items that were rated as being more memorable were actually remembered more accurately.

Finally, looking at the study time data, there was an effect of EORs on study time, $F(1,18) = 7.83, MSE = 6128913.37, p = .012$, meaning that items studied for longer were rated as being more memorable.
Table 15
Experiment 6: P(old), EORs and Study Times for High and Low Frequency Words
Median Split by EOR

<table>
<thead>
<tr>
<th>Frequency</th>
<th>p(old)</th>
<th>Study time (ms)</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low EOR</td>
<td>.79 (.04)</td>
<td>4291 (714)</td>
<td>2.92 (.24)</td>
</tr>
<tr>
<td>High EOR</td>
<td>.91 (.03)</td>
<td>5978 (1060)</td>
<td>4.67 (.22)</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low EOR</td>
<td>.82 (.03)</td>
<td>4429 (526)</td>
<td>2.49 (.22)</td>
</tr>
<tr>
<td>High EOR</td>
<td>.91 (.02)</td>
<td>5920 (993)</td>
<td>4.47 (.22)</td>
</tr>
</tbody>
</table>

Note. p(old) = proportion of items claimed as “old”. Old = Items that were seen in the study phase. New = Items that were not seen in the study phase. EOR = Estimation of Remembering. EORs are on a scale from 1 to 6. Standard errors are reported in parentheses.

Discussion

This experiment was very similar to Experiment 1, and the findings from this experiment are very similar to the findings of Experiment 1. Most interesting, the low frequency advantage on the hit portion of the word frequency mirror effect was eliminated. This finding was expected because the only change in procedure was the instruction manipulation. I expected that if this change had any effect at all, it would have an effect on the EORs and the study times, not on the hit data. As it turns out, this manipulation had no effect on any of the variables.

Next looking at the EOR data, consistent with Experiment 1, the high frequency words were rated as being more memorable than the low frequency words. Even though
participants are told that the test is a recognition memory experiment, they still claim the high frequency words as being more memorable. This suggests that participants are using the same heuristics in Experiment 6 as they did in Experiment 1. Also, similar to all of the experiments described using EORs, words that were rated as being more memorable were actually remembered better.

Another similarity between Experiment 1 and 6 was that there was no significant difference between the study times for the high and the low frequency words. Contrary to previous findings, in this experiment, words that were studied for longer were more likely to be recognized.
CHAPTER 3: GENERAL DISCUSSION

There were several consistencies and inconsistencies across the six experiments. First looking at the false alarm data, in all six recognition experiments the high frequency words were more likely to elicit a false alarm than the low frequency words. This finding suggests that any inconsistencies in the hit data are not due to item effects. Because the same words were used in all experiments, and because each word was randomly assigned to the “old” or to the “new” condition, if there were certain items that were more likely to be called “old” than other items, those effects would be just as likely to appear in the hit data as the false alarm data. Furthermore, because the high frequency words consistently elicit more false alarms than the low frequency words, but there is inconsistency between the high and the low frequency words in terms of the hits, there is a dissociation between the hits and the false alarms. This dissociation suggests that different variables mediate the hits and the false alarms. Although not hugely surprising, this finding has implications for the single process accounts of recognition memory. These implications will be discussed in detail at the end of the General Discussion.

One inconsistency across the experiments was observed in the hit portion of the mirror effect. In Experiments 1 and 6, there was no difference in the hits between the high and the low frequency words. In all other experiments the low frequency words were more likely to elicit a hit than the high frequency words. There are several possible reasons why the standard mirror effect was not observed in certain experiments.
The failure to find the hit portion of the word frequency mirror effect does not seem to be related to the self-paced training. In Experiment 3, the training was self-paced, yet a standard mirror effect was observed. The failure to observe the hit portion of the mirror effect appears to have something to do with evaluating one's memory for the item. When participants are asked to rate the pleasantness of the items (instead of how memorable they thought the item was) the standard mirror effect is observed. Similarly, as demonstrated by Experiment 4, if the EORs are based on a brief experience with the words, the standard word frequency mirror effect is observed. Thus it appears as that when participants perform EORs that are based on a rich encoding experience and that are based exclusively on evaluations of one's own memory, the hit portion of the mirror effect is eliminated.

To determine the reason that the low frequency word advantage was eliminated, the exact nature of the change must be determined. There are three possible ways in which the high frequency hits might come to equal the low frequency hits in Experiments 1 and 6. The hits of the high frequency words could have been increased to the level of the low frequency words, the hits of the low frequency words could have been decreased to the level of the low frequency words, or it could be a combination of both. To get an idea of what is going on, the overall hit rates from Experiments 1 and 6 can be compared to the overall hit rates from another comparable experiment. For this comparison, Experiment 2 was used as the baseline because the procedure of that experiment was exactly the same as Experiments 1 and 6, with the exception that no EOR was performed. As such, any differences in the hit rates must be due to the EOR. If the overall hit rates from Experiments 1 and 6 are higher than the hit rates from Experiment 2, then the hits
on the high frequency words were increased. If the overall hit rates in these experiments are lower than the hit rates in Experiment 2, that would indicated that the hits on the low frequency words were decreased. If the overall hit rates are about the same, that would indicate both an increase in the hit rates of the high frequency words and a decrease in the hit rates of the low frequency words in Experiments 1 and 6. The overall hit rate in Experiment 2 was 86%. The overall hit rates in Experiments 1 and 6 were 90% and 86% respectively. It appears as though in Experiment 1, the high frequency words elicited more hits, whereas in Experiment 6, the high frequency words elicited more hits and the low frequency words elicited fewer hits.

The next thing to determine is whether the change in the hit portion of the word frequency mirror effect came about because of changes at the time of the training or changes at the time of the test. Because all experimental manipulations were conducted in the training phase of the experiments, it seems as though the differences in the hit rates must be caused by manipulations during the training. However, it is possible that the training manipulations caused the participants to adopt different strategies during the test phase of the experiments. For example, it is possible that forcing participants to do an EOR causes them to realize that they have no idea if they will remember the items or not. This realization might cause participants to rely more on familiarity and less on recall. If this was the case, then the hits on the high frequency words would increase. However, if the changes in the hit rates come about because of different strategies during the test phase of the experiment, that effect should also be observed in the false alarm data because the false alarms are also based on familiarity. However, the false alarm data of all experiments (except for Experiment 2) are very similar. Thus, because there is no
difference in the false alarms patterns between experiments, it appears as though the
differences in the hit patterns come about because differences that arise during the
training phase.

The next question is: what is it about the training experiences of Experiments 1
and 6 that cause the hit portion of the word frequency mirror effect to be eliminated? One
explanation might be that the EOR causes a ceiling effect in the hit data. It is possible
that performing an EOR increases the likelihood that any word will be remembered, but
because the low frequency words are already recognized very well, the only items that
can benefit are the high frequency words. There are several reasons why this explanation
is unlikely. First, the low frequency words are correctly recognized about 90% of the
time, as such, there is still the possibility of increasing the probability of a hit for the low
frequency words. Second, as was discussed above, the hit rates from Experiments 1 and 6
do not differ to a great degree from the hit rates of the other experiments. However, for
the sake of completeness, to further investigate the possibility of a ceiling effect, a
between subject median split analysis was performed on the overall accuracy. To do this,
the overall accuracy for each subject was calculated, then the subjects were rank ordered
based on their overall accuracy. The half of subjects with the highest overall accuracy
was separated from the half of subjects with the lowest overall accuracy (in the case of an
odd number of participants, the median participant was excluded from analysis). The
high frequency hits were compared to the low frequency hits, but just for the half of
subjects with the lowest overall accuracy. If a ceiling effect was responsible for the
elimination of the mirror effect, then the difference between the high and the low
frequency words should still be seen in the subjects that are not performing at the ceiling.
In Experiment 1, the p-value for this difference was .572. Thus it would appear that this finding is not due to a ceiling effect.

Another possible reason that the high frequency hits equal the low frequency hits in these experiments could be that labeling an item as memorable causes that item to become memorable. It could be that when participants claim that a word is likely to be remembered, they make an added effort to ensure that it is well known. In all five experiments that investigated EORs, the high frequency words were rated as being more memorable than the low frequency words. If claiming that a word is likely to be remembered causes that word to become more memorable, and the high frequency words are rated as being more memorable than the low frequency words, then it follows that the hits for the high frequency words will increase more than the hits for the low frequency words. Some support for this idea is that items given higher EOR are remembered better than items given low EORs. This explanation does however have flaws. The main problem with this explanation is that the EOR patterns in Experiment 4 are the same as the EOR patterns in Experiment 1. If the elimination of the hit portion of the mirror effect was caused by the fact that the high frequency words are rated as being more memorable, the same effect should be seen in Experiment 4, but it is not.

Another interesting difference between Experiments 1 and 6 and Experiment 4 is the overall EORs. The overall EORs in Experiments 1 and 6 were significantly higher than the EORs in Experiment 4. This could also explain why the mirror effect disappears in Experiments 1 and 6. While in all of the experiments the items rated as being more memorable were indeed recognized better, overall the EORs are higher in Experiments 1 and 6 than in Experiment 4. As such, the “High EORs” in Experiments 1 and 6 are
higher than the “High EORs” of Experiment 4. Because the high frequency words are remembered better than the low frequency words in all experiments, the hits for the high frequency words become more memorable in Experiments 1 and 6 because the High EORs are higher than the High EORs in Experiment 4.

Yet another possible explanation of the elimination of the hit portion of the word frequency mirror effect is that factors that influence the EOR also cause the words to become more memorable. To explore this possibility it must be determined what the basis is on which participants make their EORs. In all five experiments that investigated EORs, the high frequency words were rated as being more memorable than the low frequency words. One possibility for this difference was that the participants thought the test would be a recall test, in which case the high frequency words usually are remembered better. However, this possibility seems unlikely because in Experiment 6, participants are explicitly told that the test is a recognition test, and they are told exactly what happens in a recognition test. Despite these instructions the participants still rate the high frequency words as being more memorable than the low frequency words. Perhaps a more likely explanation of this finding is that participants are using a fluency heuristic when performing their EORs.

As discussed in the introduction, it has been found that participants sometimes base other metacognitive judgments on fluency heuristics (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Matvey, Dunlosky & Guttentag, 2001; Koriat, Bjork, Shefer, & Bar, 2004). It would be expected that the high frequency words would be experienced with greater fluency because, by definition, they are more familiar to the participants.
There are, however, different types of fluency heuristics that the participants could use. Much of the research conducted that has looked at Judgments of Learning has suggested that participants use a retrieval fluency heuristic when performing their JOLs. This is very unlikely to be the case in the current experiments because participants make their EORs immediately after seeing the word, resulting in very high retrieval fluency of all the words. Another type of fluency participants could use is conceptual fluency. Conceptual fluency can be defined as how easily the concept of the word comes to mind. If this were the case, then the finding that the high frequency words are given higher EORs would also be expected because the concepts of the high frequency words would be easier to think about. Could it be that thinking about the concept of the words causes the high frequency words to be remembered better? It is possible that participants think about the concepts of the high and low frequency words in different ways in the different experiments. When participants do not have to make an EOR, it is possible that they think only about the first thing that comes to mind. In this case, because the low frequency items are experienced less, they will be more distinctive and thus easier to recognize. However, when the participants know that they will have to perform an EOR, they could try to think about how much they know about the words. This processing will not only cause the high frequency words to be given higher EORs, but it will also cause the high frequency words to be given more elaborate processing in the training phase. That is, participants can conceive of more information relating to the high frequency words that the low frequency words, which would cause the high frequency words to be easier to recognize in the test phase of the experiment. This type of evaluation would not happen in Experiment 3, because participants are no longer making a memory evaluation.
Furthermore, this type of evaluation would not happen in Experiment 4 because the EOR is based on a brief exposure to the item. As a result, the participant does not have time to think about everything they know about the item and might make their EOR based on the ease with which the concept of the word came to mind. Furthermore, in Experiment 4, the EOR is performed before the participants study the words. As such, the participants would not study the words in the same way as when they are expecting to perform an EOR.

In summary, when simply thinking about the concept of the word, the low frequency words will be easier to recognize than the high frequency words because the concept is more distinctive. However, when an EOR has to be made on the item, the participants think more elaborately about the concepts of the words, and because the participants have more experiences with the high frequency words, they elaborate more during the study time of the high frequency words.

Next looking at the study time data, in Experiment 2, the low frequency words were studied for longer than the high frequency words; however, in all other experiments there was no difference in the study times between the high and the low frequency words. This failure to find a consistent effect is not surprising considering the nature of the study times. The study times are quite variable both between and within subjects; as such, it is difficult to find any reliable effects. The differences that were significant showed that the low frequency words were studied for longer than the high frequency words. It is possible that upon encountering the items, the participants perform an EOR in order to establish how much study time to allot to each item. If the high frequency words are rated as being more memorable, then it follows that the participants will study the low
frequency items for longer. Although this is a possibility, it is difficult to make any
strong conclusions about this because only one out of the six differences was significant.

In addition, the items that were studied for a long time were not remembered
better than the words studied for a short period of time. This was the case in four of the
five experiments that examined the study times. This finding is quite interesting because
it suggests that the quantity of study at the time of training can not be responsible for the
hit portion of the standard word frequency mirror effect.

Finally, the relationship between the EORs and the study times are variable across
the six experiments in which both of these variables were measured. In all of the
experiments, there was no difference in the EORs between items studied for a long period
of time and items studied for a short period of time. However, in Experiments 2 and 6,
the items given higher EORs were studied for longer than the items given lower EORs.
While these findings are relatively simple to explain on their own, explaining all of the
findings together is not as straightforward.

First, the finding that in certain experiments the items given higher EORs are
studied for longer is simple to explain because the EORs are made after the participants
have studied the items, leading to one of two processes. One possibility is that because
participants perform their EORs after studying the items, the items studied for longer will
tend to be given higher EORs. This relationship did not hold up in Experiment 3 because
participants did ratings unrelated to memory and in Experiment 4 because participants
had not yet studied the word when they performed their EOR. What is perplexing is the
finding that there is no difference in the EORs between the items studied for a long
period of time and the items studied for a short period of time.
There are a couple of possible reasons that this finding could have come about. First is that the differences observed between the study times for items given high and low EORs could be due to outliers. However, after doing an outlier analysis in which all trials in with study times that were outside the range of 2.5 standard deviations from the mean were eliminated, the same results were obtained. Consistent differences within the fast and slow half of study times, yet otherwise evenly distributed, would lead to these results. For example, the higher EORs might happen to be situated at the high half of both the fast and the slow study times. If this was the case then you would see no difference in the EORs when median split by study time, but a difference in the study times when median split by EORs. Even though the median split procedure is simply a correlation, and causation cannot be determined, this procedure does give some information about the nature of the relationship between the two variables that a standard correlation will not give.

The Source of Activation Confusion (SAC) model proposed by Reder and colleagues (2000) argues that the mirror effect arises because of differences in how the high and low frequency words are represented in memory. The low frequency words have a higher threshold of activation resulting in a lower probability of experiencing a feeling of familiarity for those words, but they have fewer related event nodes resulting in a higher probability of recall than the high frequency words because the event nodes are more distinct from each other. Because this model is a dual process account of the mirror effect, the dissociation between the hits and the false alarms is not difficult to explain. Because the false alarms are based on familiarity and the hits are based on familiarity and recollections, the finding that the hit portion of the mirror effect can be affected without
affecting the false alarm portion is simple; it must the recollection part of the hits that are being affected.

In Experiments 1 and 6, the hit portion of the word frequency mirror effect was eliminated. As discussed, one possible reason that this occurred was that when the participants knew an EOR was coming, they thought about the EORs as they studied the words. Because the participants know more about the high frequency words, this led to more elaboration for the high frequency words eliminating the hit portion of the mirror effect. The SAC model could explain this finding by arguing that when the participants elaborated on the high frequency words, the event node made more connections with contextual nodes making the high frequency words easier to retrieve in the test phase of the experiment.

The SCAPE framework would explain the results in a similar fashion. In Experiments 1 and 6, the high frequency words were elaborated more than the low frequency words during the training phase of the experiment. This would cause a greater fit between the high frequency words and the context in which they were presented. In turn, this would cause an increase in the probability that those items would be recollected in the test phase of the experiment. The extra elaboration advantage of the high frequency words matched the distinctiveness advantage of the low frequency words.

Finally, the Attention/Likelihood theory has more difficulty accounting for the data that was observed. First, because the Attention/Likelihood model is a single process account of the mirror effect, it argues that the false alarm pattern is tied to the hit pattern. In fact, in the study conducted by Glanzer, Adams, and Iverson (1991), the fact that the false alarm portion of the effect was tied to the hit portion of the effect was considered
evidence that there was in fact only one process responsible for the effect. However, all of the experiments presented in this thesis, the false alarm pattern of the effect was constant with more false alarms being elicited by the high frequency words. This finding occurred even in experiments in which the hit portion of the effect was affected. This finding cannot be explained simply by arguing that the selection criterion of the experiment has changed. If this were the case, then a change in the low frequency words should be seen as well. One assumption of the Attention/Likelihood model is that there is no difference in the number of marked features between the high and the low frequency words before the experiment begins. By removing that assumption, and by arguing that the high frequency words have more marked features before the experiment begins, this theory can account for the data. If this were the case, then it would be expected that any test time manipulation would affect the familiarity distributions of the old words, not the new words, and as such the high frequency words should always elicit more false alarms than the low frequency words. However, this assumption would be difficult to change given the results of Glanzer, Adams, and Iverson's study described in the introduction. Essentially, they argued that the fact that the hits were tied to the false alarms was evidence in favor of the Attention/Likelihood model, and the experiments described in this thesis fail to find such a relationship.

Another theoretical problem with this model is that participants must estimate the number of marked features they expect for an item of that type. Presumably this means that participants know the difference between the high and the low frequency words. But for subjects to know the difference between the high and the low frequency words, they must be using some cue other than the number of features marked, because that is what
they are evaluating. This model must include a way in which the participants can differentiate between the high and the low frequency words.

The Attention/Likelihood model can explain the fact that the mirror effect was not observed when an EOR was performed because the number of marked features for the high frequency words might change. If participants are able to elaborate on the high frequency words more when they perform an EOR, then it follows that more features will become marked. If this is the case, then the high frequency words will have just as many marked features as the low frequency words and the hit rates will even out.

Overall, it seems as though the dual process accounts of the mirror effect have an easier time explaining the data found in these experiments. The single process account can account for the data, but only when small changes in the assumptions have been made. However, changing the assumptions of the model then changes the explanation for previous experiments and it remains to be seen if the model can withstand these changes.

**Concluding Remarks**

Perhaps the most interesting and theoretically important finding presented in this thesis is the finding that the metacognitive judgments can affect the memory for the items that are being evaluated. It has been argued that in a delayed JOL condition, the JOL procedure actually causes items given higher JOLs to be remembered better (Spellman and Bjork, 1992, Kimball and Metcalfe, 2003). However, this is the first time an immediate metacognitive judgment has been found to affect the ultimate memory for the item. This finding is interesting because it suggests that metacognitive judgments are not completely independent of the processes that they are evaluating. An analogy to this situation would be trying to measure how fast your lawn is growing by measuring the
trimmings from the mower each week. However, it is possible that mowing the lawn actually causes the lawn to grow faster, thus by using that procedure, you never have a true estimate of how fast the lawn is growing independent of the mowing. This is perhaps the same problem that is occurring with the use of metacognitive measures. I am certainly not suggesting that metacognitive measures should not be used. On the contrary, I think that metacognitive psychology is an important field to examine. I am suggesting that researchers should use care in determining how their metacognitive measure is affecting what the metacognitive measure is intended to evaluate.

An interesting line of research to pursue might be in determining the exact processes by which the metamemory evaluations affect memory. While the experiments in this thesis suggest possibilities, there are certainly some unexplored ways in which the metamemory evaluations can influence memory. Furthermore, there may be other areas of cognition in which this effect might be observed. Claiming something is easy to perceive might make it easier to perceive, and claiming something is difficult to pay attention to might cause that items to become more difficult to pay attention to. There is an entire unexplored research line that can be dedicated to the question of how metacognition affects cognition. This line of research would be important because not only would it provide information about how metacognition occurred, but how also various cognitive processes occur.
REFERENCES


