UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

CAN RECOLLECTION EXPLAIN MIRROR EFFECTS IN ITEM RECOGNITION? EVIDENCE FROM HUMANS AND A COMPUTATIONAL MODEL

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

by

TROY ANTHONY SMITH

Norman, Oklahoma

2010

CAN RECOLLECTION EXPLAIN MIRROR EFFECTS IN ITEM RECOGNITION?
EVIDENCE FROM HUMANS AND A COMPUTATIONAL MODEL

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

_____
Dr. Daniel R. Kimball, Chair

_____
Dr. Scott D. Gronlund

_____
Dr. Rickey P. Thomas

_____
Dr. Joseph Lee Rodgers, III

_____
Dr. Marlys Lipe

Acknowledgements

I would like to thank my faculty mentor, Dr. Daniel Kimball, for the guidance, training, and opportunities he has given me over the last 5+ years of graduate school. From the time I started as volunteer in his lab, Dan has consistently encouraged me to think deeply and critically about the cognitive processes that underlie human behavior. His training in theoretical analysis and experimental design has been an invaluable part of my graduate education, but Dan's emphasis on actually doing research—not just learning about it from textbooks and classroom exercises—has probably had the largest impact on my development as a professional scientist. Before I was officially even a first year grad student, Dan gave me the opportunity to work with him on the development of the fSAM recall model. Many of the ideas in this thesis grew directly from that experience.

I would also like to thank the other members of my committee and the rest of the faculty, staff, and graduate students in the OU Department of Psychology. Despite the fact that I transferred to OU late in my graduate studies, you welcomed me into the department and made me feel at home. Even though I hail from south of the Red River, I am proud to be a Sooner.

Table Of Contents

List of Tables

List of Figures

Abstract

This thesis examines the role of recollection in the word frequency mirror effect. In two experiments, participants studied lists comprised of sets of associatively related words and sets of unrelated words from 4 levels of word frequency. Following a short distractor task, participants took an item recognition test with confidence ratings (Experiment 1) or old-new judgments followed by remember-familiar judgments (Experiment 2). In Experiment 1, the standard word frequency mirror effect for comparisons of LF words to HF words was observed, but the effect did not obtain for comparisons of MF, HF, or VHF words. In Experiment 2, a complete word frequency mirror effect was observed, and the patterns for hit rates and remember judgments for targets were almost identical. These findings run counter to predictions from Bayesian likelihood models (Glanzer, Hilford, & Maloney, 2009) but are consistent with the hypothesis that mirror effects are the result of differences in recollectability between stimulus classes (Joordens & Hockley, 2000). Attempts to develop a computational process model to account for mirror effects and boundary conditions on those effects such as those observed in Experiment 1 are also discussed.

Chapter 1

Introduction

Prior to the 1990's, models of recognition memory generally assumed that performance on recognition tasks is a function of the strength or familiarity of items in memory (e.g., Anderson & Bower, 1972, 1973, 1974; Gillund & Shiffrin, 1984; Hintzman, 1988). Although these models used different architectures and operationalized memory strength in dramatically different ways, they made one common prediction: All other things being equal, manipulations that increase the strength of the representation of stimuli in memory will increase the probability that old items will be recognized as old (i.e., the hit rate, HR) and the probability of new items being recognized as old (i.e., the false alarm rate, FAR). Thus, when Glanzer and Adams (1985) showed that a wide variety of manipulations result in a very different pattern—a mirror effect wherein recognition memory performance is improved by simultaneously *increasing* the HR and *decreasing* the FAR—memory modelers began to question the validity of the assumptions behind strength-based theories of recognition memory.

In particular, Glanzer and his colleagues have argued that the regularity with which mirror effects are observed requires a reconceptualization of how recognition memory operates (Glanzer, Adams, Iverson, & Kim, 1993; Glanzer, Hilford, & Maloney, 2009), and although mirror effects have received relatively little attention in the development of general recognition theories, they have become one of the phenomena that are "at the heart of testing and evaluating" computational models of

recognition memory (Ratcliff & McKoon, 2000, pg 575). This emphasis on mirror effects has led many memory model developers to reject strength-based global memory models (GMMs) of recognition such as the search of associative memory model (SAM; Gillund & Shiffrin, 1984) in favor of models that use Bayesian likelihood ratios rather than memory strength as the basis for making recognition decisions. One of the key advantages of these newer models is that they naturally produce mirror effects as a byproduct of the Bayesian decision process (e.g., Dennis & Humphreys, 2001; Glanzer et al., 1993; Glanzer et al., 2009; Shiffrin & Steyvers, 1997).

However, there is a growing body of evidence showing that there are boundary conditions on mirror effects and that these effects are not as regular as has been generally assumed (for a review see Greene, 2007). These findings imply that the current emphasis on explaining mirror effects is incomplete and that models of recognition memory should be able to account for the important cases in which mirror effects do not occur, as well as the cases in which they do. Classic single-process, strength-based recognition models predict the absence of mirror effects but do not generally predict their presence (Glanzer & Adams, 1985; but see Gillund & Shiffrin, 1984). On the other hand, models that use a Bayesian likelihood transformation easily predict the presence of mirror effects but do not generally predict their absence (Glanzer et al., 2009). Thus, none of the extant models are sufficient.

There were two primary goals for this research project. First, I wanted to investigate the effects of normative word frequency and semantic association on true and false recognition. A large number of studies have shown mirror effects when comparing memory for low frequency and high frequency words; however, Estes and

Maddox (2002) showed that these effects are moderated by the magnitude of the disparity in relative frequency. There are reasons to believe that semantic association may also moderate the word frequency mirror effect, and it is not known what effects word frequency has on associatively induced false recognition. To examine this issue I designed and ran two novel experiments that combined the word frequency paradigm used by Estes and Maddox with a variant of the Deese-Roegider-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995) as used by Kimball, Muntean, and Smith (2010).

Second, I wanted to develop a computational model of recognition memory that could parsimoniously account for mirror effects and their absence without using Bayesian likelihood transformations. To this end, I modified the fSAM recall model (Kimball, Smith, & Kahana, 2007) by incorporating processes that assess memory strength based on familiarity and recollection and use the results of those processes to make recognition decisions. I tested a number of possible model variants in a series of simulations that examined the ability of the model to account for 1) the shape of ROC and *z*-ROC curves relating hits and false alarms as a function of subjective confidence in item recognition and 2) word frequency mirror effects.

Thus, the research project encompassed two distinct phases, and the organization of this thesis reflects those phases. The first half of this thesis focuses on the mirror effect, including the two new experiments I conducted. First, Chapter 2 discusses the basics of measuring recognition memory performance and reviews the major theories of recognition memory. Chapter 3 then reviews mirror effects in recognition memory, including the evidence for viewing mirror effects as a regularity of

memory, the theoretical explanations that have been offered to explain mirror effects, and some reported boundary conditions for mirror effects. The results from two unique experiments that were performed to investigate the effects of word frequency and semantic association on true and false recognition are reported in Chapters 4 and 5. These experiments were also designed to provide data to guide and constrain the development of the fSAM recognition model.

The second half of this paper focuses on the development and testing of a series of computational models of recognition based on the fSAM recall model. Chapter 6 starts by briefly reviewing extant computational models of recognition memory, focusing on how these models deal with two critical issues in recognition memory theory—the shape of ROC curves and the word frequency mirror effect. The chapter then discusses issues related to model development and testing, including the rationale for my modeling approach, and briefly describes of the SAM framework and the fSAM model of recall. Chapter 7 covers the development of the fSAM recognition model through two sets of simulations. The chapter describes a number of possible ways in which familiarity and recollection could be implemented within the fSAM framework, reports the results of a test of the ability of 36 different model variants to account for the shape of ROC curves in episodic recognition, and shows that none of these models are able to generate a word frequency mirror effect. Finally, Chapter 8 presents concluding remarks, including discussions of why the fSAM model cannot account for mirror effects, what characteristics a strength-based model would need in order to account for mirror effects, and possible directions for future research.

Chapter 2

Fundamentals of Recognition Memory

In the standard recognition memory experiment, participants study a list of items such as words or pictures, wait for some amount of time, and are then tested on their memory for those items. There are a number of ways in which the recognition test can be given. The simplest type of test involves old-new judgments: Participants are presented with a series of items, some of which were in the previously studied list and some of which were not, and are asked to determine whether each item is "old" (i.e., was in the list) or "new" (i.e., was not in the list). A common variation on this approach is to ask participants to rate their confidence that each test item was a previously studied item using a Likert scale that ranges from "sure old" to "sure new." These confidence ratings can be converted to old-new judgments or used to build receiver operating characteristic (ROC) curves as described below. Another variation is the remember-know procedure in which participants are asked to make an additional judgment for each item that they identify as old, indicating their phenomenological experience regarding the item, such as whether they can recall specific details about the item's presentation during study or just have a feeling that the item was studied (Gardiner & Java, 1991; Rajaram, 1993; Tulving, 1985).

Measuring Recognition Memory Performance

Participants' performance on the recognition test can be measured in a number of different ways, depending on the method that is used to administer the recognition test. For purposes of scoring, items that were previously studied are referred to as

targets, and items that were not previously studied are referred to as *lures*. For old-new judgments, correct identification of a target as old is termed a *hit*, and incorrect identification of a lure as old is termed a *false alarm*. Conversely, incorrect identification of a target as new is a *miss*, and correct identification as new is a *correct rejection*. Because these sets of scores are complementary, old-new recognition memory performance is usually measured by the proportion of hits (hit rate, HR) and the proportion of false alarms (false alarm rate, FAR).

Signal detection theory. Recognition memory performance can be described in more detail, though, using the measurement tools of signal detection theory (SDT). As applied to memory performance, SDT assumes that the strength of items in memory is distributed in a Gaussian fashion, with different item classes having different distributions, as shown in Figure 1. Prior to being studied, all items are assumed to be part of the unstudied or lure distribution. Studying items strengthens them. Classical signal detection theory assumes that all studied items are strengthened equally so that the target distribution is shifted to the right relative to the lure distribution but the

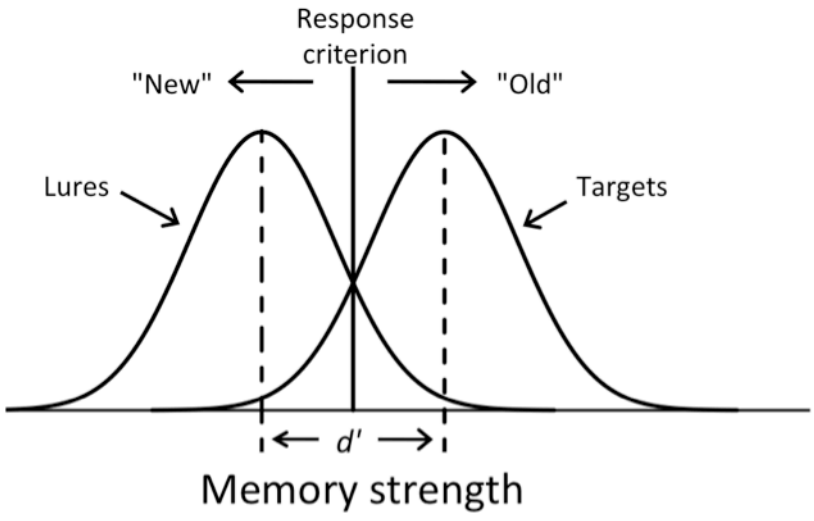

Figure 1. Distribution of memory strengths in an equal-variance signal detection model.

variances of the distributions remain equal, as shown in Figure 1 (Wickens, 2002). However, if there is variability in the strengthening of items during study, both the mean and the variance of the target distribution increase, yielding an unequal variance model (Wixted, 2007a).

Participants are assumed to make recognition memory decisions by comparing the strength of a memory probe to a response criterion. If the strength is above the criterion, then the item is judged to be "old"; otherwise the item is judged to be "new." Thus, studied items whose strength is above the threshold result in hits, while unstudied items that are above the criterion result in false alarms. Based on these assumptions, SDT provides a number of ways to measure recognition performance, including metrics for the discriminability between the distributions, response bias, and the shapes of the distributions (for reviews, see Macmillan, 2002, and Wickens, 2002).

Discriminability. The most basic metric in SDT—discriminability or sensitivity (*d'*)—measures the standardized distance between the means of the distributions. If the underlying distributions are assumed to have equal variances (σ), then *d'* can be easily calculated as (HR – FAR) / σ. When the equal variance assumption is relaxed, the more general formula $d' = z(\text{HR}) - z(\text{FAR})$ is preferred (Macmillan & Creelman, 1991; Wickens, 2002). Because *d'* is a standardized measure, in theory it allows researchers to compare recognition memory performance across conditions that differ on a number of dimensions—including the type and number of stimuli, instructional manipulations, and subject population—and a vast majority of recognition memory studies use *d'* as the measure of sensitivity. The main drawback to using single-point measures of sensitivity such as *d'* is that unless the equal variance assumption is met, these measures

are subject to systematic bias effects that can make meaningful theoretical interpretations virtually impossible (Verde, Macmillan, & Rotello, 2006). Given the considerable evidence that there are usually significant differences in the distributions of the strength of items in memory (Wixted & Stretch, 2004), these bias effects may be particularly problematic for studies of recognition memory performance.

Receiver operating characteristics (ROCs). Fortunately, SDT provides a tool for measuring recognition memory performance that takes into account not only the relative locations of the distributions, but also the variances and response biases. Isosensitivity curves, more commonly called receiver operating characteristics (ROCs), show the change in performance across different criteria. Because they show the change in sensitivity over all possible response biases, ROCs are much more powerful than single-point metrics such as *d'* (Macmillan & Creelman, 1991).

ROCs can be constructed from theoretical distributions based on hypotheses about the operation of memory processes or from data obtained in empirical studies with human participants (Wickens, 2002). To construct a theoretical ROC, the researcher needs to specify the parameters for the means and the variances of the target and lure distributions, along with a set of criterion levels. These criterion points represent different levels of response bias, and are often mapped to confidence levels (e.g., Yonelinas, 1994). Once these parameters have been specified, the cumulative probabilities for hits and false alarms at each level of response bias are calculated, and these are plotted against each other to produce an ROC plot in probability space (see Figure 2, Panel A). If the hits and false alarms are normalized prior to being plotted, a *z*-space ROC (*z*-ROC) can be produced (see Figure 2, Panel B). Empirical ROCs are

created by measuring subjects' performance at different levels of response bias, such as by asking them to make confidence judgments instead of binary old-new judgments, and then plotting the resulting cumulative hit and false alarm rates across the levels of response bias (for a brief tutorial on constructing empirical ROCs, see Yonelinas & Parks, 2007). As with theoretical *z*-ROCs, the cumulative hit and false alarm rates can be normalized and plotted in *z*-space.

ROC plots are particularly useful because they show a picture of the changes in sensitivity across different levels of bias that can be used to infer the shapes of the underlying distributions (Macmillan & Creelman, 1991; Wickens, 2002). In probability-space ROCs, the diagonal line from bottom-left to top-right represents the case where sensitivity is zero (i.e., discrimination is at chance, $d' = 0$). Parallel lines above the diagonal represent increasing levels of sensitivity. Assuming the distributions are Gaussian with equal variances, if the sensitivity across confidence levels is constant, then the points will lie on a symmetric curve with an intercept of zero. Distortions such as a non-zero intercept or a non-symmetric curve (see Figure 2, Panel A) indicate a

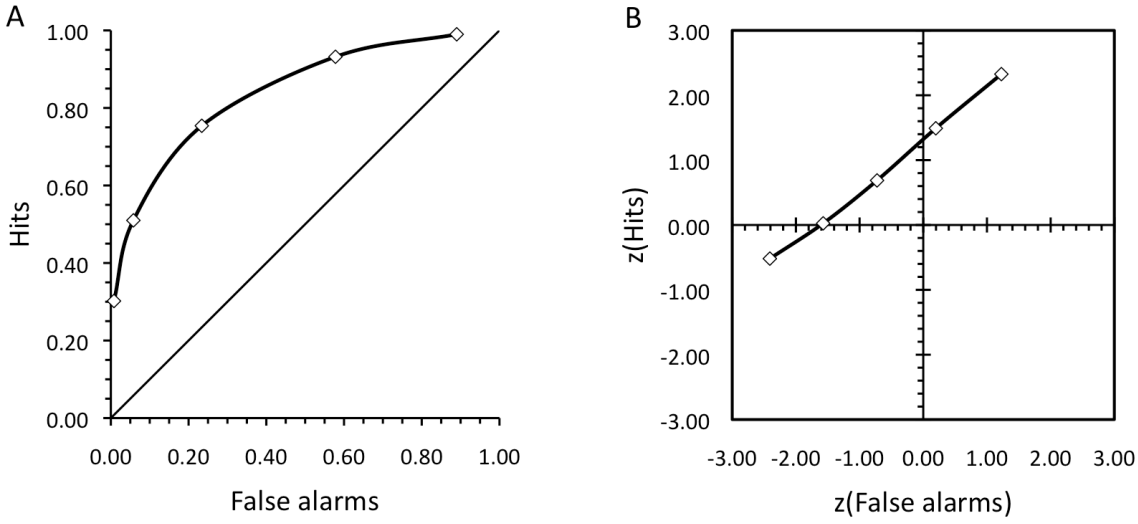

Figure 2. Typical ROC (Panel A) and z-ROC (Panel B) for item recognition memory.

change in sensitivity across confidence levels. These changes are produced when the distributions have unequal variances (e.g., Wixted, 2007a) or when an additional threshold-based decision process is added to the SDT decision process (e.g., Yonelinas, 1994).

Although distortions can make the standard probability-space ROC difficult to interpret, when transformed into *z*-space the resulting *z*-ROC (see Figure 2, Panel B) can easily be interpreted in terms of its linearity, slope, and intercept (Wickens, 2002). A linear *z*-ROC indicates that the underlying distributions are Gaussian, while a non-linear *z*-ROC is indicative of non-Gaussian distributions. The slope of the *z*-ROC measures the ratio of the variances of the target distribution to the variance of the lure distribution. Thus, a linear *z*-ROC with slope of 1.0 represents constant sensitivity based on Gaussian target and lure distributions with equal variances. A slope greater than one indicates that the lure distribution has a larger variance than the target distribution, and a slope less than one indicates that the target distribution has a larger variance. Finally, for linear and near linear *z*-ROCs, the intercept measures the overall sensitivity (*d*' or d_a).

Estimating the contributions of familiarity and recollection. The dual-process signal detection (DPSD) model (Yonelinas, 1994, 1997) interprets distortions in the ROC as evidence of the operation of two distinct memory processes—familiarity and recollection. The DPSD theory is discussed in detail later in this chapter, but for the moment, I note that DPSD provides a measurement model that can be fit to empirically obtained data and thereby used to estimate the relative contributions of familiarity and recollection processes to participants' recognition memory performance.

An Excel 2007 spreadsheet that uses the Solver tool to fit the DPSD model by minimizing the sum of squared errors is available from the Human Memory Lab at the University of California, Davis (http://psychology.ucdavis.edu/labs/Yonelinas/Software.html).

Whereas the SDT measurement techniques discussed above were initially designed to assess performance on perceptual tasks and only later adapted to measure recognition memory performance (Wickens, 2002), Tulving's (1985) remember-know procedure and Jacoby's (1991) process-dissociation procedure were designed specifically to assess the relative contributions of the underlying cognitive processes to that performance. Both of these procedures assume that two independent processes are used to make recognition decisions and that the relative influence of these processes can be measured, either directly or indirectly.

The remember-know procedure. In the recognition version of the remember-know procedure (e.g., Gardiner & Java, 1991), participants are given an old-new recognition test and for each item they rate as old, they are asked to indicate whether they identified that item as a previously studied item because they were able to consciously recollect the previous encounter with the item (a "remember" response) or because it just seemed to be old (a "know" response). The key assumption underlying the remember-know procedure is that these two responses tap into different memory processes or systems. Tulving (1985) originally devised the procedure to assess the relative use of episodic memory (remember response) and semantic memory (know response), but the procedure is most often used to assess the relative contributions of recollection and familiarity processes. And although some of the underlying

assumptions have been questioned (e.g., Rotello, Macmillan, Reeder, & Wong, 2005; Wais, Mickes, & Wixted, 2008), the remember-know procedure remains an important tool for investigating the cognitive processes used in recognition and evaluating models of recognition memory (Cohen, Rotello, & Macmillan, 2008; Yonelinas & Parks, 2007).

Process-dissociation procedures. Rather than directly measuring the contribution of cognitive processes through introspective reporting, Jacoby's (1991) process-dissociation procedure attempts to estimate their use with a logical subtraction approach. In this procedure, participants study a list of items in which the perceptual characteristics and/or processing tasks vary across the words. For example, some words may be presented auditorily while others are presented visually. After an appropriate delay, participants are given either an exclusion test or an inclusion test. On the exclusion test participants are instructed to identify as old only the items that were presented in a particular manner (e.g., only the words that were heard but not those that were read). It is assumed that participants have to use a recollection process in order to reject items from the non-target category (e.g., the read items), so that the mistaken recognition of items from the non-target category is due to a failure of recollection. By subtracting the probability of making an old judgment for the non-target items on the exclusion test from the probability of responding old to those same items on the inclusion test, the researcher can obtain an estimate of the extent to which familiarity processes were used. This can then be used to estimate the use of recollection processes. A number of studies using the process-dissociation method have shown that familiarity and recollection estimates can be reliably dissociated using experimental manipulations such as varying the study duration or using divided attention during study

(for a review, see Kelly & Jacoby, 2000). However, the viability of the process dissociation procedure is heavily reliant on the assumption that the underlying processes are independent (Curran & Hintzman, 1995), and there are boundary conditions on its use (Jacoby, 1998).

General Theories of Recognition Memory

Single-process or dual-process? Traditionally, recognition memory theories have been divided into single-process theories and dual-process theories, depending on how many memory processes people are assumed to use when deciding whether a stimulus has been encountered previously (Yonelinas, 2002). As the name implies, single-process theories assume that only one memory process is required in order to make a recognition decision. Early single-process theories, called threshold models, assumed a discrete process akin to free recall that could be explained using state diagrams (e.g., Atkinson, 1963; Luce, 1963). More recent single-process theories, including those incorporated into the global matching memory models (GMMs) that are discussed in Chapter 6, assume that the strength of items in memory can be indexed on a single continuous dimension and that recognition decisions are made using a decision process that is described by classical SDT (e.g. Gillund & Shiffrin, 1984).

On the other hand, dual-process theories assume that recognition decisions involve two different memory processes and that recognition is the result of the additive effects of these two processes (Mandler, 1980). The first of these processes is a feeling of "knowing" that you have encountered this stimulus before. Various theories ascribe this sense of familiarity as being due the overall strength of the memory representation for that stimulus (Mandler, 1980; Yonelinas, 1994) or to the ease of processing the test

stimulus (Jacoby, 1991). The second process, usually called recollection, is a recall-like process in which one tries to retrieve contextual information about previous encounters with the stimulus. For clarity in this paper, unless I indicate otherwise, I use the terms *familiarity* to refer to strength-based memory processes and *recollection* to refer to recall-like processes, without endorsing any particular theoretical definition of these terms.

The traditional division into single- and dual-process theories was useful because there was a clear distinction between the two classes of theories. Single-process theories postulated a single memory process that fed into a single decision process. Conversely, dual-process theories postulated two memory processes, each with its own associated decision process. However, the development of new hybrid recognition theories has blurred this distinction to a point where it may no longer be a meaningful way to distinguish between the theories. There are at least three reasons why the single- vs. dual-process distinction needs to be rethought.

First, the constructs associated with the cognitive processes involved in recognition are not always defined in the same way across the different classes of theories (Yonelinas, 2002). A related problem is that some of the critical constructs that are shared across the different classes of recognition theories are ill-defined, making it difficult to create consistent operational definitions for measuring the constructs. For example, with the notable exception of computational models such as the SAM model (Gillund & Shiffrin, 1984), theories based on SDT postulate the existence of distributions of memory strength, but they often do not define the construct of "memory strength" or describe how these distributions arise (T. Smith & Kimball, in press). This

ambiguity has lead to a great deal of debate in the literature as to whether the familiarity distributions for targets and lures (an unobservable construct) have equal variances or unequal variances, and much of the single- vs. dual-process controversy had been fueled by this issue (see, e.g., Parks & Yonelinas, 2007; Wixted, 2007a; Wixted 2007b; Yonelinas & Parks, 2007).

Second, as mentioned above, it is not always clear which types of processes are being described by the terms "single-process" and "dual-process." In particular, the term "single-process" has been used to describe theories that postulate a single *memory process* as well as theories that postulate a single *decision process*. For example, because the unequal variances signal detection (UVSD) model argues that recognition performance can be described with a SDT model that involves only one decision process, some researchers classify it as a "single-process" model (e.g., Diana, Reder, Arndt, & Park, 2006; Slotnick & Dodson, 2005). But if one allows for the possibility that the memory strength that is used in the UVSD decision process is derived by combining sources of evidence generated by familiarity and recollection memory processes, then it can also be classified as a dual-process or hybrid model (Wixted, 2007a; Wixted & Stretch, 2004). As Heathcote, Raymond, and Dunn (2006) have pointed out, this is true in general: Most so-called "single-process" recognition theories could be more properly described as multiple-process or hybrid theories because they implicitly assume that the memory strength used in the signal detection process is a composite of strengths from multiple memory processes. Multidimensional SDT-based theories such as the STREAK model (Rotello, Macmillan, & Reeder, 2004) that assume bivariate distributions of memory strengths from different sources along with different

criteria for each dimension further blur the distinction between single- and dual-process theories. Should these theories be classified as "single-process" because they use a single SDT process for making the recognition decision? Or should they be classified as "dual-process" because they include two different memory processes?

Third, a consensus that a simple single-process model—be it a threshold model or a single dimensional strength-based familiarity model—is insufficient to explain even the basic phenomena in recognition memory has clearly been reached (Diana et al., 2006; Rotello et al., 2004; Slotnick & Dodson, 2005; Wixted, 2007a; Yonelinas & Parks, 2007). Importantly, a consensus that two distinct memory processes—generally termed "familiarity" and "recollection"—drive recognition performance also appears to be forming. Today, the major debate seems to be over the nature of these processes (e.g., is recollection continuous or discrete) and the manner in which they are combined to make recognition decisions (e.g., are familiarity and recollection combined into a single composite strength upon which a SDT decision process operates or does recollection drive a separate threshold decision process?). The major theories that address these issues are described in more detail below.

Mandler's dual-process theory. Mandler (1969, 1980) popularized the idea that recognition decisions can be made using two distinct processes and introduced a mathematical model to describe how these processes interact. Mandler defined familiarity as a sense that a particular stimulus (the test stimulus) has been previously encountered and proposed that this is a measure of the extent to which the test stimulus has been integrated with other items and contexts. That is, familiarity is a global measure of the strength of the item in memory, which may or may not be accompanied

by the ability to identify contextual details about previous occurrences of that item. According to Mandler, familiarity is usually accompanied by a recall-like attempt to retrieve contextual details about previous occurrences of that item. In early versions of Mandler's (1969) model, the recognition decision process was assumed to be serial: Familiarity is first assessed, and then a memory search is engaged, if necessary. In later versions of the model (e.g., Mandler, 1980), the processes are assumed to run in parallel, with familiarity assessments being faster than memory search. Thus, according to Mandler, recognition decisions can be made based on familiarity, retrieval of contextual details, or a combination of both processes.

The Jacoby model. Jacoby and colleagues (Jacoby, 1991; Kelley & Jacoby, 2000) extended Mandler's dual-process model by reconceptualizing familiarity and by introducing the term *recollection* to describe the retrieval of contextual details. In the Jacoby model, familiarity is based on an assessment of processing fluency rather than memory strength. Because the processing fluency of an item increases with prior experience, the mind can infer the likelihood that an item has been previously experienced from how easy it is to process that item. This inference does not require any conscious effort and is assumed to occur automatically. Recollection in the Jacoby model is similar to the retrieval process in the Mandler model in that it is an active search of memory for contextual details and other information that was encoded during study. Thus, in the Jacoby model, familiarity is an automatic process and recollection is an effortful, conscious process. However, recollection and familiarity are still assumed to be parallel processes, with the automatic familiarity processes being faster than the conscious recollection process.

The dual-process signal detection (DPSD) model. Yonelinas (1994) integrated the Mandler (1980) and Jacoby (1991) models with signal detection theory to develop the DPSD model. In the 16 years since its introduction, the DPSD model has proven to have a high degree of explanatory power and has been influential in both cognitive psychology and cognitive neuroscience (Yonelinas & Parks, 2007). One of the key pieces of evidence for the DPSD theory is the shape of recognition memory ROCs (Yonelinas, 1994). Specifically, for item recognition, probability-space ROCs are almost universally non-linear and asymmetric, as shown in Figure 1, Panel A. These characteristics lead to *z*-ROCs that are typically linear with a slope less than 1.0 (usually around 0.8), as shown in Figure 1, Panel B (Yonelinas & Parks, 2007). This does not accord with predictions from the classical equal variance SDT model that *z*-ROCs should be linear with a unit slope. However, Yonelinas (1994) showed that adding a threshold recollection process to an equal-variance SDT familiarity process allows the DPSD model to capture the asymmetry in ROCs.

Like the Mandler (1980) model, the DPSD model assumes that recognition decisions can be based on an assessment of global memory strength (familiarity) or on the results of an active memory search (recollection), and that these processes run in parallel. Unlike previous models, though, DPSD assumes that under normal circumstances the recognition decision is made in two stages, with recollection-based decisions taking precedence over familiarity-based decisions (Yonelinas, 1994). That is, if the active memory search is successful, the item is recollected and is identified as an old item with a high confidence rating. If recollection fails or the decision needs to be made before the search terminates (e.g., as in speeded recognition tasks), the

recognition decision is made using a standard SDT decision process by comparing a familiarity value drawn from a Gaussian distribution to a single criterion to make an old-new judgment or to a set of confidence level criteria to make confidence judgments (Yonelinas, 1994; Yonelinas & Parks, 2007). The DPSD model can also accommodate remember-know judgments by assuming items that are recollected receive remember judgments and items that are judged as old based on familiarity receive know judgments (Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998).

The "dual-process" models such as Mandler's (1980), Jacoby's (1991), and Yonelinas' (1994) models all assume that two distinct memory processes, each with its own decision process, can be used to make recognition decisions; however, this assumption has been criticized as being unparsimonious (Slotnick, Klein, Dodson, & Shimamura, 2000; Wixted, 2007a; Wixted & Stretch, 2004). In principle, a recognition theory that does not require a separate recollection decision process would be more parsimonious than those that do. I next turn to one such theory that has proven to be a viable alternative to DPSD.

The unequal variance signal detection (UVSD) model. It has long been known that a single-process equal variance signal detection model is incapable of generating ROCs that look anything like the asymmetrical ROCs that are obtained in recognition memory experiments with human subjects (Green, 1960). However, a signal detection model that assumes *unequal* variances in the distributions of memory strength can fit empirical ROCs just as well, and sometimes better, than the DPSD model, even though the model lacks a threshold recollection process (Wixted, 2007a, 2007b; but see Parks & Yonelinas, 2007). Despite this fact, the DPSD model has often

been preferred over an unequal variance signal detection (UVSD) model because, until recently, there was no principled explanation of why the target distribution should have a greater variance than the lure distribution (Yonelinas & Parks, 2007; Wixted, 2007a). This problem was solved by Wixted and Stretch (2004; see also DeCarlo, 2007) by relaxing the traditional assumption that the memory strength distributions arise from a single memory source.

In Wixted's (2007a) UVSD model, memory strength is a combination of evidence from a familiarity process—as in traditional single-process SDT and the DPSD model—and a recollection process. Like the DPSD model, the UVSD model assumes that familiarity is a continuous measurement of "global" memory strength. However, whereas the DPSD model assumes that recollection is a discrete process that is separate from the SDT process, the UVSD model assumes that recollection is a continuous measurement of "item-specific" memory strength and that the two types of memory strength (familiarity and recollection) sum to an overall memory strength. This combined strength is then used to make a recognition decision as in other SDT models. Notably, although Wixted uses the terms "item-specific" and "global" strength to describe recollection and familiarity, respectively, these constructs are not clearly defined in the UVSD model. I address this point further in Chapter 7 when I describe my attempt to implement a UVSD-based process model of recognition.

The unequal variances in target and lure distributions in Wixted's (2007a) UVSD model are the result of adding the local memory strength from recollection to the global memory strength from familiarity. Unstudied items are assumed to have low recollective strengths with little variability among these values. Studied items are

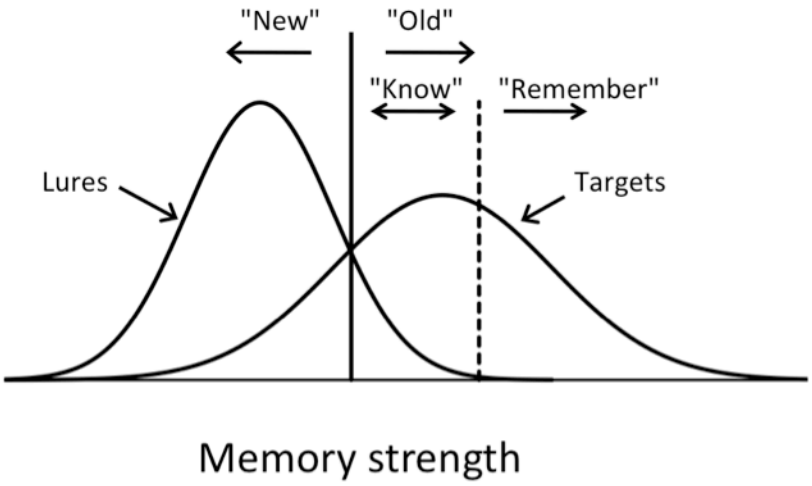

Figure 3. Distribution of memory strengths for targets and lures along with criteria for making old-new and remember-know judgments in the UVSD model.

assumed to have higher recollective strengths, but these strengths are highly variable due to differences in attention, rehearsal, and other aspects of the encoding process. When the familiarity and recollection strengths are summed, the result is a pair of normal distributions in which the unstudied items have a low mean with low variability and the studied items have a higher mean with higher variability, as shown in Figure 3.

Participants are assumed to use these combined unidimensional, unequal-variance distributions to make old-new decisions, confidence ratings, and remember-know judgments using a signal detection process, as described earlier. In the UVSD model, the critical difference between old-new and remember-know judgments is the criteria, not the processes: Old-new decisions (and confidence ratings) use one set of criteria, and remember-know decisions use another (see Figure 3). In both cases, the UVSD model predicts curvilinear, asymmetrical ROCs and linear *z*-ROCs with a slope less than 1.0 as a consequence of the differences in variance between the target and lure distributions (Wixted, 2007a).

Limitations. One important limitation of general recognition theories such as those outlined above is that the theories focus on how theoretical constructs such as memory strength, familiarity, and recollection are used to make recognition decisions but they do not detail the actual operation of the lower-level processes that give rise to memory strength, familiarity, and recollection (Yonelinas, 2002). In other words, these theories specify the operation of decision processes but require ad hoc assumptions regarding the operation of the basic underlying cognitive processes in order to be complete. As I discuss in Chapter 6, computational models such as fSAM are important tools that can be used to address this issue.

Chapter 3

Mirror Effects in Item Recognition

The mirror effect is a somewhat paradoxical effect that can occur when classes of stimuli that differ in initial strength are studied or when an experimental manipulation results in one subset of a class being strengthened more during study than another. As discussed in Chapter 2, strength-based accounts of recognition memory such as classic SDT and single-process global matching models assume that studying items strengthens them, thereby shifting the distribution of target strengths to the right, as shown in Figure 4 Panel B. If two classes of stimuli that differ in initial strength (Weak and Strong in Figure 4) are studied, the distributions of both classes should be shifted right. This ordering of distributions implies the concordant effect shown in Figure 4 Panel A in which the hit and false alarm rates for the strong class are higher than those for the weak class (Glanzer & Adams, 1990; Glanzer et al., 1993). However, when the mirror effect obtains, recognition memory performance for ostensibly stronger stimuli is improved relative to the ostensibly weaker stimuli by simultaneously increasing the hit rate and decreasing the false alarm rate, as shown in Figure 5 Panel A (Glanzer & Adams, 1985). The mirror effect is highly problematic for strength-based signal detection accounts of recognition because it implies the ordering of memory strength distribution shown in Figure 5 Panel B in which the unstudied item distributions are reversed. That is, lures for the ostensibly strong stimulus class have to be weaker than lures for the ostensibly weak stimulus class (Glanzer & Adams, 1985; Glanzer et al., 1993).

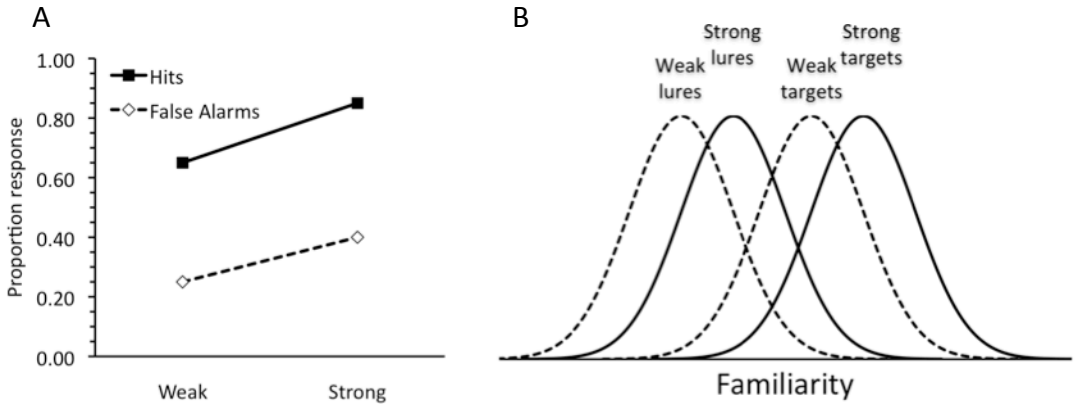

Figure 4. Example of a concordant effect (Panel A) and the strength distributions that give rise to the effect when using a signal detection decision process (Panel B).

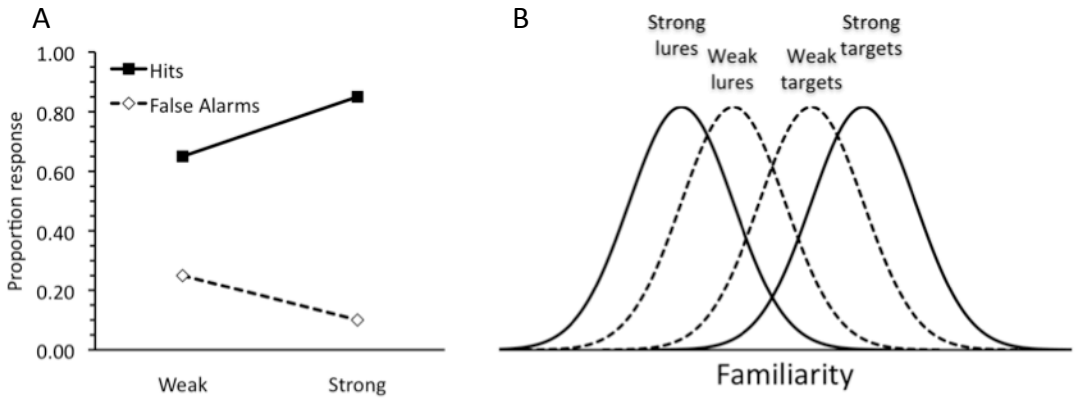

Figure 5. Example of a mirror effect (Panel A) and the strength distributions that give rise to the effect when using a signal detection decision process (Panel B).

Mirror Effects as a "Regularity" of Recognition Memory

In their seminal review paper, Glanzer and Adams (1985) present fairly compelling evidence that mirror effects occur with surprising regularity in recognition memory. They reviewed studies that required participants to perform a common recognition task such as make old-new judgments, rate their confidence, 2AFC and multiple choice and that manipulated a number of variables within subjects—including normative word frequency, concreteness or imageability, meaningfulness or familiarity of the stimuli, and pictures vs. words. For word frequency, 23 of the 24 reviewed studies showed the presence of mirror effects. For concreteness or imageability, 8 of 9 showed mirror effects. For meaningfulness or familiarity, 9 of 13 showed the effect. For pictures vs. words, 6 of 8 showed mirror effects, and for the miscellaneous variables 17 of 26 showed this effect. The fact that none of these proportions are likely to have occurred by chance demonstrates that the mirror effect is real and that it seems to occur in a vast majority of recognition studies using a wide range of variables and procedures.

Later experiments appear to strengthen these findings. In a series of five experiments, Glanzer and Adams (1990) found mirror effects for stimuli that varied in normative word frequency and concreteness and for a manipulation in which participants had to read the word backwards. Based on their previous review and these new findings, they concluded that "[a]ny variable that affects recognition accuracy…will produce the effect" (pg 12). Glanzer, Adams, and Iverson (1991) found that manipulating retention interval produces a mirror effect when using 2AFC. Hilford, Glanzer, and Kim (1997) observed mirror effects for levels of processing operations, lexical decisions tasks, and study repetition with a 2AFC test. Although the

above studies were all conducted by Glanzer and his colleagues, a number of studies by investigators not directly affiliated with Glanzer have observed mirror effects, typically with word frequency but also with other variables (e.g., Heathcote, Ditton, & Mitchell, 2006; Malmberg, Steyvers, Stephens, & Shiffrin, 2002; Reder et al., 2000).

Boundary Conditions on Mirror Effects

The fact that mirror effects occur with such resounding regularity has lead many researchers, spearheaded by Glanzer and colleagues, to view mirror effects as a general principle of recognition memory that should guide the development and testing of theories (Glanzer & Adams, 1985, 1990; Glanzer et al., 1993; Glanzer et al., 2009; Ratcliff & McKoon, 2000). Given the near incontrovertible evidence for the occurrence of mirror effects, it seems quite reasonable to expect recognition memory theories to be able to explain why mirror effects arise. However, contrary to Glanzer and Adams' (1990) prediction, variables that affect recognition accuracy do not always produce a mirror effect. Failures to find a mirror effect might be dismissed as null effects that do not require any explanation, but a number of manipulations that reliably produce concordant effects instead of mirror effects have been identified (for a review, see Greene, 2007). This suggests that there are boundary conditions on mirror effects that must be considered in the development and testing of theories. In other words, theories should be tested for their ability to explain both the occurrence of mirror effects and the occurrence of concordant patterns (or non-mirror effects). Before describing theories of mirror effects in more detail, I briefly review the evidence that mirror effects are more limited than Glanzer and Adams' (1990) rule predicts.

First, even in Glanzer and Adams' (1985) review article, several of the studies they reviewed did not show mirror effects. Again, while it might be tempting to dismiss these studies as null effects, most of these studies showed significant simple effects that were in the same direction (i.e., a concordant pattern) rather than in different directions (i.e., a mirror pattern). If any variable that affects recognition memory performance should produce mirror effects, as proposed by Glanzer and Adams (1990), why then did these studies show the opposite pattern? Admittedly, this is a weak argument against the generality of mirror effects, but it is one that has yet to be fully addressed.

A stronger argument is the fact that evidence for reliable within-subjects mirror effects with variables other than those that involve stimulus characteristics such as word frequency is inconsistent (Greene, 2007; for a review see Diana et al., 2006). For example, Stretch and Wixted (1998b) showed that when item strength is manipulated across lists, a strength-based mirror effect obtains; but when strength is manipulated within lists, the mirror effect does not obtain. Cary and Reder (2003) observed mirror effects when they varied the number of study trials within subjects, but Tussing and Greene (2001) found concordant effects with a similar manipulation.

Another factor to consider is whether the variable is manipulated within subjects or between subjects. Glanzer and Adams (1985) specifically exempted between subjects manipulations from their review because mirror effects in those studies could be trivially explained by criterion differences between the groups. Nevertheless, some of the more recent studies that have purported to show mirror effects have used a between subjects design (e.g., Hilford et al., 1997). While these studies might be interesting for other reasons, they clearly cannot be used as evidence against criterion-shift

interpretations of the mirror effect. In fact, even when variables are manipulated within subjects but across lists, mirror effects can be easily explained within a signal detection framework by allowing for differences in criteria between the lists (Stretch & Wixted, 1998a, 1998b).

Thus, only studies that manipulate the critical variable within list can truly be said to test strength-based theories of recognition (Dobbins & Kroll, 2005; Greene, 2007; Maddox & Estes, 1997; Stretch & Wixted, 1998b). This effectively eliminates many of the critical variables that have historically been used to generate mirror effects because they simply cannot be manipulated within lists, but it does leave one key variable that can be manipulated within list and is generally thought to give rise to consistent mirror effects—word frequency.

The Word Frequency Mirror Effect

Word frequency is a measure of how often a given word is used in everyday speech or in written texts and has been called "the most important variable in research on word processing and memory" (Brysbaert & New, 2009, pg 977). One of the reasons that word frequency is of interest is that it has different impacts on recall than on recognition (Gregg, 1976). Subjects typically recall high frequency words at higher rates than low frequency words, but they are more accurate at recognizing low frequency words. This interaction between recall and recognition has been dubbed the *word-frequency effect*, and has played a key roll in the testing of computational models of memory (Clark, 1992). With regard to theories of recognition memory, word frequency is of particular interest because it is one of the few variables that has been

shown to consistently generate mirror effects even when manipulated within lists (Glanzer & Adams, 1985; Stretch & Wixted, 1998b).

The word frequency mirror effect (WFME) refers to the finding that low frequency (LF) words elicit both higher hit rates and lower false alarm rates than do high frequency (HF) words (Glanzer & Adams, 1985). This finding is extremely robust. As mentioned earlier, 23 of the 24 studies of word frequency effects reviewed by Glanzer and Adams (1985) showed the presence of mirror effects. The only study that did not show a mirror effect was Shepard (1967), but this study was limited by a small sample size (n = 17) and the presence of ceiling effects. A large number of more recent studies that have compared memory for low- and high-frequency words have also shown the WFME (Arndt & Reder, 2002; Cary & Reder, 2003; de Zubicaray, McMahon, Eastburn, Finnigan, & Humphreys, 2005; Glanzer & Adams, 1990; Hockley, 1994; Malmberg & Murnane, 2002; Stretch & Wixted, 1998b).

Despite the overwhelming evidence for a WFME, there is still some question as to its generality. Specifically, all of the above studies are limited by the fact that they only examined two levels of frequency—low and high frequency. Other studies that used a broader range of frequency have found that there are clear boundary conditions to the WFME (e.g., Estes & Maddox, 2002; Heathcote, Ditton, & Mitchell, 2006; Wixted, 1992), and studies comparing memory performance for words and pseudo-words consistently show concordant patterns rather than mirror effects (Greene, 2007).

Estes and Maddox (2002) is perhaps the best example of a study that observed boundary effects for the WFME. Estes and Maddox showed that the disparity in relative frequency is a critical moderator variable for observing mirror effects with

word-frequency. In two experiments that used five levels of word frequency, including non-words and very low frequency words, the mirror effect only obtained for comparisons between very high frequency words and low frequency words—the standard manipulation in word-frequency effect studies. Concordant patterns were observed for all other comparisons of interest, including those between very high frequency words and very low frequency words and those between high frequency words and low frequency words. Re-analyses of data from other studies that used more than two levels of word frequency also showed these patterns. The Estes and Maddox study is discussed in more detail in Chapter 4.

Theoretical Responses to Mirror Effects

Mirror effects have had a major impact on the development and testing of computational models of recognition memory. These models are reviewed in more detail in Chapter 6, but for the current discussion two important points need to be noted. First, because the global memory models that use memory strength to make recognition decisions cannot account for mirror effects, many researchers have written off these models as having been falsified (e.g., Diana et al., 2006). Second, in keeping with the logic of single-process theories, researchers have developed a new class of computational memory models that use Bayesian likelihood transformations as the basis for recognition decisions (e.g., Glanzer et al., 1993; McClelland & Chappell, 1999; Shiffrin & Steyvers, 1997). As a consequence of the Bayesian likelihood transformation, these models predict that mirror effects will obtain for any set of stimuli where one stimulus class is stronger than the other (Glanzer et al., 2009). As discussed

in Chapter 6, though, these models do not seem to be able to account for cases when there are clear differences in stimulus strengths and mirror effects do not obtain.

By contrast, mirror effects have played a surprisingly small role in the single- vs. dual-process debate described earlier. In fact, with one notable exception, I have not been able to find any discussion of mirror effects in the literature on the general recognition memory theories reviewed earlier in Chapter 2, including the DPSD and UVSD models. Wixted and Stretch (2004) address strength-based mirror effects based on criterion shifts and use the UVSD model to predict that such effects should extend from old-new judgments to remember-know judgments. However, they do not address word frequency effects, nor do they describe how the UVSD model could accommodate mirror effects in situations where a criterion shift is unlikely to occur. A few researchers have argued that mirror effects support a dual-process account of recognition memory (e.g., Cary & Reder, 2003; Joordens & Hockley, 2000; Reder et al., 2000), but these arguments have had almost no impact on the broader debate. Nevertheless, this point deserves a closer look.

Recollection as an Explanation for Mirror Effects

Recollection and the WFME. The Source of Activation Confusion (SAC) model developed by Reder and colleagues (Reder et al., 2000) **i**ntegrates the concepts of familiarity and recollection from the Mandler (1980) and Jacoby (1991) dual-process theories in a computational model that can be used to simulate human performance in various recognition memory tasks, including old-new recognition and remember-know judgments. The SAC model, loosely based on the spreading activation model of Collins and Loftus (1975), represents memory as a collection of interconnected item nodes and

context, or *event*, nodes. During study, the item nodes are activated and links between item nodes and event nodes are formed or strengthened. As in other spreading activation theories, activation decays over time. In the SAC model, familiarity is based on the strength of activation of the stimulus during test. Thus, familiarity is a function of how often and how recently the test stimulus has been encountered. Recollection is a search for a particular event node that is associated to the test stimulus (i.e., the node for the study episode) and is based on the fan from the test stimulus to the event nodes. When tested, items that are connected to only a few event nodes will have a high probability of increasing the activation of the critical event node above a threshold, thereby generating a recollection experience. These items can be said to be highly recollectible. Conversely, items that are connected to a large number of event nodes will be less likely to activate any one of them above the threshold and will therefore be less recollectible.

Reder et al. (2000) showed that the combination of these two processes allows the SAC model to reproduce the WFME when appropriate assumptions are made for representing low- and high-frequency words in the model. First, they assumed that the baseline activation level of the item nodes is a function of the word's normative frequency. Because familiarity is based on the strength of the item node, this assumption causes unstudied high-frequency lures to have higher familiarity values—and therefore higher false alarm rates—than low-frequency lures. Second, they assumed that the fan was also a function of normative word frequency, such that high frequency words were connected to more event nodes than were low frequency words. This assumption implements the idea that high-frequency items have been experienced in a greater variety of contexts than low-frequency words and are therefore less

recollectible. The high recollectability of low-frequency words causes higher hit rates for low-frequency targets than for high frequency targets. In summary, the SAC model uses a two-factor account in which the WFME is due to differences in both the familiarity and the recollectability of low- and high-frequency words.

Recollection as a general explanation of mirror effects. Joordens and Hockley (2000) generalized the two-factor account to explain other causes of mirror effects and, importantly, cases wherein mirror effects do not occur. According to Joordens and Hockley, the stimulus classes in mirror effect experiments have different levels of preexperimental familiarity. Consistent with the assumptions of classical SDT and the GMMs, studying target items from either stimulus class increases their familiarity equally, shifting the target distributions to the right, as shown in Figure 4. When judgments are made on the basis of familiarity, the ordering of the distributions causes subjects to respond "old" more often to items from the stimulus class with a higher preexperimental familiarity. Because judgments for unstudied items (lures) are based predominantly—if not exclusively—on familiarity, this implies that the low-familiarity stimulus class should always exhibit a lower false alarm rate than the high-familiarity stimulus class. If judgments for target items are also based on familiarity, a concordant pattern in which hit rates and false alarm rates are higher for the more familiar stimulus class than for the less familiar stimulus class should be observed, as predicted by classical SDT and GMMs (Glanzer et al., 1993).

However, within the dual-process framework, judgments for studied items can also be based on recollection. If, as Joordens and Hockley (2000) argue, items in the high-familiarity class have been associated to a greater number of contexts in the past

than items from the low-familiarity class, then recollection strengths for the targets will mirror familiarity strengths. That is, the low-familiarity class will be more recollectible than the high-familiarity class. Thus, when subjects are able to use recollection, the increased recollectability of the less familiar stimulus class will drive up the hit rate for that class, thereby generating a mirror effect.

Of course, under normal circumstances, subjects are unlikely to use the same process for every target item on the recognition test. When the combined contributions of familiarity and recollection are considered, a more complex pattern arises in which the results can range from a concordant effect to a null effect to a mirror effect depending on the relative contributions of the two processes to recognition decisions. In the next two chapters, I present the results from two experiments designed to examine the relative impacts of recollection and familiarity on mirror effects in order to test this hypothesis.

Chapter 4

Experiment 1

Within the community of recognition memory researchers, there are three major perspectives on mirror effects. Glanzer and colleagues argue that mirror effects are ubiquitous and should be one of the key phenomena that are used to test recognition memory theories (Glanzer & Adams, 1985; Glanzer et al., 1993; Glanzer et al., 2009). They have further argued, rather convincingly, that strength-based familiarity models of memory including classic SDT models and GMMs cannot account for mirror effects in general terms and are therefore inadequate explanations of recognition memory processes. In keeping with the "single-process" logic that underlies those models, Glanzer et al. (1993) proposed that recognition memory decisions are made using a Bayesian process rather than a direct assessment of memory strength. This assumption naturally produces mirror effects and has been built into a number of computational models of recognition memory (e.g., Glanzer et al., 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997; see Chapter 6 for more details). These Bayesian likelihood models predict that, in general, any variable that affects recollection performance should produce a mirror effect (Glanzer et al., 2009). Throughout the remainder of this thesis, I refer to this idea as the *Bayesian likelihood hypothesis*.

In contrast to the view advocated by Glanzer, some researchers have argued that although mirror effects may be ubiquitous, they are not all that important because they can be explained within the signal detection framework by allowing for criterion shifts (e.g., Wixted, 1992; Wixted & Stretch, 2004). However, this is not a very strong

argument: Even proponents of this view concede that within-list mirror effects cannot be readily explained by criterion shifts (Stretch & Wixted, 1998a, 1998b). Thus, a more general explanation of mirror effects is needed.

A third group of researchers agree with Glanzer that mirror effects are important, but point out that the failure to find mirror effects is just as, if not more, important (e.g., Hintzman, Caulton, & Curran, 1994; Joordens & Hockley, 2000; Diana et al., 2006). Working within a dual-process framework, this group has argued that mirror effects are caused by differences in recollection and that failures to find mirror effects are due to the selective use of familiarity. Throughout the remainder of this paper, I refer to this idea as the *recollection hypothesis*.

Overview of the Experiments

There were two goals for Experiments 1 and 2. First, these experiments were designed to test competing predictions from the Bayesian likelihood hypothesis and the recollection hypothesis by manipulating word frequency across a wider range than is typically done (cf. Estes & Maddox, 2002) and by crossing the word frequency manipulation with semantic relatedness. Second, these experiments were designed to provide data to constrain the development of a computational model of recognition memory based on the recollection hypothesis as discussed in Chapters 6 and 8.

I chose to manipulate word frequency because the Bayesian likelihood hypothesis predicts that differences in word frequency should always produce mirror effects (Glanzer et al., 1993; Shiffrin & Steyvers, 1997) whereas the recollection hypothesis suggests that word frequency should only produce a mirror effect if it leads to differences in recollection (Joordens & Hockley, 2000; Reder et al., 2000). When

recognition memory for unrelated low- and high-frequency words is compared, a mirror effect almost universally obtains (for reviews, see Glanzer & Adams, 1985; Joordens & Hockley, 2000). However, in the three studies that have varied word frequency over a broader range, the mirror effect obtained only for comparisons of low-frequency words to high- or very high-frequency words (Estes & Maddox, 2002; Rao & Proctor, 1984; Wixted, 1992).

I crossed word frequency with semantic relatedness for three reasons. First, Monaco, Abbott, and Kahana (2007) recently showed that a neural network using semantic associations from word association space (WAS; Steyvers, Shiffrin, & Nelson, 2004) could account for the WFME, although this model was quite limited in its scope and required assumptions that may be somewhat questionable, as I discuss later (see Chapter 8). The finding that "word frequency is encoded in the semantic structure of language" (Monaco et al., pg 204) suggests that semantic associations play an important role in the WFME, and that there might be an interaction between semantic relatedness and word frequency effects in recognition memory. The use of WAS to account for the WFME is also important for the development of the fSAM recognition model described later in Chapter 7.

Second, the Bayesian likelihood hypothesis and the recollection hypothesis both predict that there should be a mirror effect for semantic relatedness, albeit for different reasons. A number of studies have shown that semantic relatedness impacts recognition performance, with recognition accuracy being higher for semantically related words than for semantically unrelated words (Neely & Tse, 2007). Because semantic relatedness clearly affects recognition performance in this way, the Bayesian likelihood

hypothesis predicts a mirror effect for this variable. Semantic relatedness is also thought to play an important role in recollection (Brainerd, Reyna, Wright, Mojardin, 2003; Brainerd, Wright, Reyna, & Mojardin, 2001). Brainerd et al. (2001, Experiment 3) showed that when subjects studied mixed lists with a relatively even distribution of semantically related and semantically unrelated words, the semantically related words were more recollectable than were the semantically unrelated words. If this holds for the stimuli used in Experiments 1 and 2, as is reasonable to assume, then the recollection hypothesis predicts a semantic relatedness mirror effect with hit rates being higher and false alarm rates lower for semantically related lists than for semantically unrelated lists. In Brainerd et al. (2001), the advantage for semantically related lists came at the cost of elevated false alarms for critical lures associated to the semantically related lists, a phenomenon that Brainerd et al. called "phantom recollection."

By using lists of semantically associated words in Experiments 1 and 2, I was also able to look for possible word frequency effects in the DRM false memory paradigm. Only one published study to date has examined possible effects of word frequency on false recognition, but the results from that study are ambiguous. Anaki, Faran, Ben-Shalom, and Henik (2005) used the mirror effect to test whether false recognition of an unstudied critical lure was due either to activation of the critical lure during study of its semantic associates as hypothesized by activation-monitoring theory (Roediger, Balota, & Watson, 2001), or to the use of gist traces during test as hypothesized by fuzzy trace theory (Brainerd et al., 2001). According to Anaki et al., if the locus of false recognition is during study, a word frequency effect should be observed for the critical lures; conversely, if the locus is at test, then a word frequency

effect should not occur. In their first experiment, Anaki et al. manipulated the normative familiarity (as a proxy for word frequency) of the list words and critical lures along with the strength of association from the critical lure to the list words (i.e., backward associative strength). They observed the standard WFME, and for lists with high backward associative strength the pattern of false alarms to the critical lures was similar to the pattern of hits for the studied words, supporting the activation account. However, there were a number of potential confounds in their design, including contamination of recognition by a prior recall test. In their second experiment they corrected some of these confounds, but they did not observe a reliable mirror effect. The experiments I report below used a different design than Anaki et al., thereby avoiding the confounds in their study while still enabling me to investigate the effect of word frequency on semantically induced false recognition.

Basic design. The basic design for the experiments combines elements of two experiments from completely different paradigms. The manipulation of word frequency was inspired by an experiment designed to investigate the influence of familiarity on the WFME (Estes & Maddox, 2002, Experiment 2). The manipulation of semantic relatedness, including the use of semantically related lures during test, is based on an experiment performed by Kimball et al. (2010) to investigate the effects of spreading semantic activation during a recognition test on false memory.

Word frequency. Estes and Maddox (2002, Experiment 2) made two major changes to the standard paradigm used to investigate the WFME. Instead of using only 2 levels of word frequency, Estes and Maddox used 4 levels of word frequency along with a non-word condition. This gave them a total of 5 lexical conditions—non-words,

very low frequency words, low frequency words, high frequency words, and very high frequency words. At the beginning of the experiment, they presented subjects with the targets and the lures from a given lexical condition from 0 to 8 times in order to increase the familiarity of the test stimuli. During this familiarization phase, the subjects were not told which items would be targets and which would be lures. The subjects were then given a subset of the stimuli to study. Following the study task they were given a recognition test with confidence judgments over the studied and unstudied (but familiarized) words.

In order to examine possible effects of recollection on the WFME, I borrowed two aspects of Estes and Maddox's (2002) design. First, I manipulated word frequency over a wider range than is typically done. However, because non-words and very low-frequency words that subjects are unlikely to have ever encountered are arguably qualitatively different types of stimuli than are words that subjects have encountered prior to the experiment, I chose to replace the non-word and very low-frequency conditions with a medium frequency condition. Second, rather than just collecting old-new judgments, I had participants make confidence judgments on the recognition test in Experiment 1 and remember-familiar judgments in Experiment 2. I did not use the stimulus pre-exposure procedure.

Semantic association and false memory. Kimball et al. (2010, Experiment 1) had participants study lists of 40 words that were constructed either from thematically related or thematically unrelated sets of words. The thematically related sets were the 36 DRM lists from Stadler, Roediger, and McDermott (1999) that are commonly used to investigate semantically induced false memory. These DRM lists are constructed

such that all of the words in each list are semantically related to a single word that is not studied; this word is referred to as the *critical lure*. After studying each 40-item list, subjects were given a recognition test over that list. When the studied list had used thematically related words, the lures on the test were drawn either from sets of unstudied thematic word lists or from a set of unstudied, unrelated words. When the studied list had used unrelated words, the lures on the recognition test were drawn from a set of unstudied thematic word lists. Semantically induced false memory was tested by including the critical lures for both studied and unstudied thematic lists near the end of the recognition test. This design allowed Kimball et al. (2010) to examine the influence of prior study and prior test on false recognition of the critical lure.

In order to examine possible mirror effects due to semantic relatedness and word frequency on both true and false memory, I borrowed three components from Kimball et al.'s (2010, Experiment 1) design. First, in addition to the sets of unrelated words that have been used in most other WFME studies, I included sets of semantically related words in the study lists. Because the standard DRM lists do not control for word frequency, I created a custom set of stimuli as described in the Methods section below. Accordingly, the targets on the recognition test were drawn from both associative and non-associative sets of studied words. Second, the lures on the recognition test included words from unstudied associative lists as well as unstudied non-associative lists. Third, the critical lures from both studied and unstudied (but tested) associative lists were included near the end of the recognition test.

Method

Participants. Participants were 72 undergraduate students enrolled in psychology courses at the University of Oklahoma who participated for partial course credit. All participants spoke and read English fluently.

Materials and design. All stimuli were selected from the University of Southern Florida word association norms (USF norms; Nelson, McEvoy, & Schreiber, 2004) and the SUBTLEXus word frequency norms (Brysbaert & New, 2009). The USF norms provide an estimate of the semantic association between pairs of words by giving people a target word and asking them to report the first word that comes to mind in response to that target word. When responses are tallied over large numbers of words and individuals, the forward associative strength (from the target to the response) and the backward associative strength (from the response to the target) can be calculated for many pairs of words (Nelson et al., 2004). The SUBTLEXus word frequency norms are a new set of norms that are superior to the often-used Kučera and Francis (1967) norms because they are derived from a larger corpus that is more representative of the natural use of contemporary American English, thereby providing a more accurate measurement of word frequency (Brysbaert & New, 2009). The SUBTLEXus norms include over 74,000 unique words (including proper nouns) and provide a measurement of how often each of these words occurs in American English using several different metrics. For this experiment, stimuli were selected using the Lg10WF metric (base 10 logarithm of the number of times the word occurred in the SUBTLEXus corpus of 51 million words), but to facilitate comparison to other studies frequency is reported using the $SUBLT_{WF}$ metric (word frequency per million words).

A total of 48 six-item word lists were constructed using words that occurred in both the USF and SUBTLEXus norms. These lists were divided into sets based on the following 4 levels of word frequency: low frequency (LF), medium frequency (MF), high frequency (HF), and very-high frequency (VHF). The range and mean word frequency for each level, along with comparisons to other WFME experiments, are shown in Table 1. As can be seen from the table, there is considerable variation in how word frequency has been operationalized in previous studies, and the operational definitions used in this experiment overlap and extend the ranges used in other studies. Note that the word frequency for the HF condition in other experiments tends to fall around the MF range for this Experiment.

For each level of word frequency, 6 associative lists and 6 matched non-associative lists were constructed. The associative lists were similar to DRM lists that are commonly used to study false memory (e.g., Roediger, Watson, McDermott, & Gallo, 2001): Each list consisted of a set of words within the specified frequency range that are all semantically associated to a single critical word as measured by backward associative strength in the USF norms (Nelson et al., 2004). The non-associative lists were constructed by randomly selecting words from the specified frequency range without regard to semantic association. Finally, an additional 24 non-associated low-to-medium frequency words (3.5 to 47.1 occurrences per million words, $M = 14.5$) were selected to act as primacy and recency buffers. The associative word lists and their critical lures, the matched non-associative lists, and the words used for primacy and recency buffers are listed in the Appendix.

Table 1. Operational definitions of word frequency in Experiments 1 and 2 and in other selected studies of word frequency mirror effects in recognition.

		Word frequency (per million words)		
Study	Condition	Min	Max	Mean
Experiments 1 and 2				
	LF	1.9	2.8	2.3
	MF	49.4	77.1	60.3
	HF	202.6	487.2	306.7
	VHF	627.2	9773.4	2293.1
Estes and Maddox (2001)				
	VLF	< 1 in 6 million		--
	LF	~1	~1	--
	HF	2.0	39.0	--
	VHF	41.0	2714.0	--
Glanzer and Adams (1990, Experiment 2)				
	LF	--	--	12.2
	HF	--	--	164.0
Higham et al. (2009)				
	LF	5	7	--
	HF	500	--	--
Kim and Glanzer (1993)				
	LF	0.0	8.0	--
	HF	40.0	--	--
Malmberg and Murnane (2002)				
	LF	1.0	10.0	--
	HF	50.0	--	--
Reder et al (2001)				
	LF	--	--	1.6
	HF	--	--	142.0
Stretch and Wixted (1998)				
	LF	0.0	3.0	1.6
	HF	40.0	--	98.9

The experiment used a completely within-subjects 2 (Semantic Relatedness: associated vs. non-associated) × 4 (Word Frequency: LF, MF, HF, VHF) design with all variables manipulated within-list. The master set of stimuli described above was used to create 3 sets of study-test lists for each subject and repeated measures were taken across the resulting sets of study-test cycles. The assignment of specific lists to study and test trials was randomly determined for each subject. A desktop computer with the E-Prime 2.0 software package was used to present the stimuli, control the timing of tasks, and record participant's responses.

Procedure. After informed consent was obtained, participants were given oral instructions by the experimenter. Participants were told that they would be memorizing lists of words and solving math problems and that they should do their best on both tasks. The experimenter then provided step-by-step instructions for each task as samples of the displays corresponding to each step were presented on the computer. As part of the instructions, participants were given a short practice session consisting of a 10-item study list and a 10-item recognition test.

The experiment proper consisted of 3 study-test cycles. On each of these cycles, each participant studied a list of 56 words, performed a one-minute distractor task, and then took an item recognition test. During study, words were presented one at a time for 2.5s each with a 500ms interstimulus interval. The study list consisted of 4 primacy buffer words, 48 words from associative and non-associative lists for each normative frequency presented in random order, and 4 recency buffer words (see Figure 6). After the last study word was presented, participants were given a set of arithmetic problems to solve for 60s before the item recognition test began. As shown in Figure 6, the

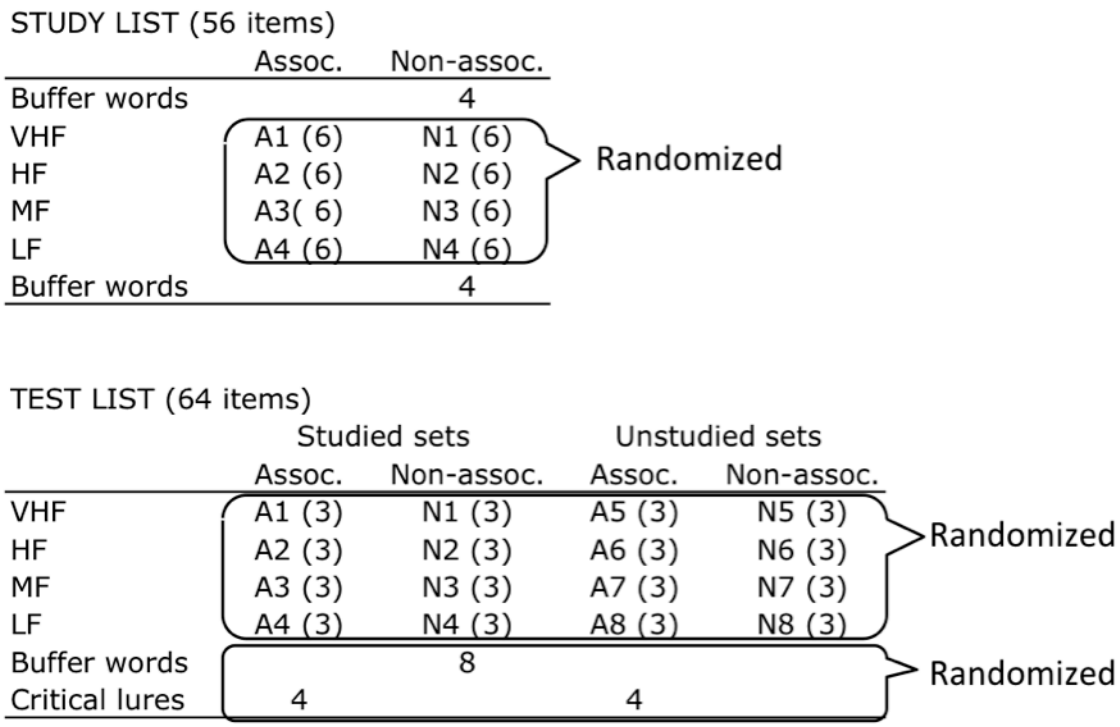

Figure 6. Composition of a study-test list set in Experiments 1 and 2. A_x = associative stimuli set *x* and N_y = non-associative stimuli set *y*, where *x* and *y* are indices that identify unique stimulus sets.

recognition test consisted of half the studied words from each frequency/association sub-list (24 targets) and a corresponding set of unstudied words (24 lures) presented in random order, followed by a final sequence comprising the 8 buffer items, the 4 critical lures from the studied associative lists, and the 4 critical lures from the unstudied associative lists, also presented in random order. Participants were asked to make old-new confidence judgments for each test item by indicating their confidence on a 1-6 scale, where 1 represented highly confident that the item was studied and 6 represented highly confident that the item was not studied.

Results and Discussion

I conducted two separate sets of analyses for Experiment 1. In the first set of analyses, I converted participant's responses to old-new judgments by scoring

confidence ratings of 1, 2, or 3 as "old" and confidence ratings of 4, 5, or 6 as "new". I then analyzed these old-new judgments for effects of normative word frequency, semantic association, and possible interactions between word frequency and semantic association. The second set of analyses uses participants' confidence ratings to examine ROC curves using the DPSD model for evidence of changes in recollection and familiarity as a function of word frequency and semantic associations.

Old-new judgments: Effects of word frequency. The first three rows of Table 1 show the mean proportions of items endorsed as old for targets (hits), non-critical lures (false alarms), and critical lures (critical lure FAs) as a function of word frequency, collapsed across semantic association. A 3 (Item Type: target, non-critical lure, critical lure) x 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA revealed significant main effects of item type, $F(2, 142) = 349.72$, $MSE = 0.066$, $p < .001$, and word frequency, $F(3, 213) = 32.45$, $MSE = 0.019$, $p < .001$, but these main effects were qualified by a significant interaction, $F(6, 426) = 29.72$, $MSE = 0.015$, $p <. 001$. A set of planned contrasts revealed that this interaction was due to a pattern consistent with the presence of an overall word frequency mirror effect, with hit rates decreasing monotonically as normative frequency increased, $t(71) = -3.95$, $SEM = 0.066$, $p <.001$, while false alarms to non-critical lures and critical lures both increased, $t(71) = 11.02$, $SEM = 0.061$, $p < .001$, and $t(71) = 8.39$, $SEM = 0.091$, $p < .001$, respectively.

However, there was a qualitative difference in the relationship between word frequency and old judgments for targets and lures. Specifically, there was a significant non-linear component for hit rates, $t(71) = -2.53$, $SEM = 0.023$, $p = .014$, such that low frequency targets were judged to be old more often than were targets of moderate

Table 2. Mean proportions of items endorsed as old for targets (hits), non-critical lures (false alarms), and critical lures (critical lure FAs) as a function of word frequency and semantic relatedness in Experiment 1.

	Word frequency									
	Low		Medium		High		Very High			
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Overall										
Hits	0.84	(0.016)	0.78	(0.018)	0.76	(0.017)	0.76	(0.018)	0.78	(0.009)
False alarms	0.14	(0.018)	0.24	(0.021)	0.29	(0.024)	0.35	(0.023)	0.25	(0.012)
Critical lure FAs	0.26	(0.024)	0.28	(0.023)	0.33	(0.024)	0.50	(0.029)	0.34	(0.014)
Associative Lists										
Hits	0.85	(0.019)	0.80	(0.021)	0.76	(0.020)	0.77	(0.023)	0.80	(0.011)
False alarms	0.14	(0.018)	0.23	(0.025)	0.26	(0.026)	0.37	(0.027)	0.25	(0.013)
Non-associative Lists										
Hits	0.82	(0.017)	0.76	(0.020)	0.75	(0.020)	0.75	(0.022)	0.77	(0.010)
False alarms	0.15	(0.020)	0.24	(0.023)	0.31	(0.028)	0.32	(0.026)	0.26	(0.013)
Critical lure FAs										
Studied lists	0.26	(0.032)	0.31	(0.031)	0.40	(0.035)	0.54	(0.035)	0.38	(0.018)
Unstudied lists	0.25	(0.029)	0.26	(0.027)	0.26	(0.033)	0.45	(0.040)	0.31	(0.017)

frequency, $t(71) = 3.49$, $SEM = 0.017$, $p < .001$, high frequency, $t(71) = 4.41$, $SEM = 0.018$, $p < .001$, or very high frequency, $t(71) = 3.91$, $SEM = 0.020$, $p < .001$, but there were no significant differences in hit rates between MF, HF, and VHF targets. Importantly, there was no hint of such nonlinearity for non-critical lures, $t(71) = 1.51$, $SEM = 0.024$, $p > .10$. This suggests that the cognitive mechanisms responsible for the overall word frequency mirror effect may behave differently for studied items versus unstudied items.

There was also a qualitative difference in the effects of word frequency for non-critical lures versus critical lures. Whereas word frequency increased false alarms to non-critical lures linearly, there was a significant quadratic component for the effect of word frequency on false alarms to critical lures, $t(71) = 3.40$, $SEM = 0.040$, $p = .001$, that was in the opposite direction from the pattern observed for targets. There was no significant difference in the probabilities of endorsing critical lures from low and moderate frequency associative lists, $t(71) = 1.18$, $SEM = 0.024$, $p > .10$; there was only a marginally significant difference between false alarms to critical lures from moderate versus high frequency lists, $t(71) = 1.86$, $SEM = 0.025$, $p = .068$; but there was a large difference between false alarms to critical lures from high versus very high frequency lists, $t(71) = 5.70$, $SEM = 0.029$, $p < .001$. One possible explanation for this pattern is that the VHF associative word sets are highly confusable—that is, less distinct—relative to the other associative word sets, thus impairing participants' ability to effectively use source monitoring processes to reject the critical lure.

Old-new judgments: Effects of semantic relatedness and word frequency. To test for effects of semantic association and possible interactions between semantic

association and word frequency, I conducted a separate set of analyses on just the targets and the non-critical lures. The means for these analyses are shown in Table 1 (rows 4-7). A 2 (Item Type: target vs. non-critical lure) × 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA revealed significant main effects of item type, $F(1, 71) = 415.35$, $MSE = 0.195$, $p < .001$, and word frequency, $F(3, 213) = 8.93$, $MSE = 0.023$, $p < .001$ on participants' willingness to judge an item as old. Overall, semantic relatedness did not have a significant effect on old judgments, $F(1, 71) = 1.52$, $MSE = 0.019$, $p > .10$.

However, these main effects were qualified by the presence of significant two-way interactions. Item type interacted with semantic relatedness, $F(1, 71) = 4.48$, $MSE = 0.017$, $p = .038$, and with word frequency, $F(3, 213) = 56.16$, $MSE = 0.020$, $p < .001$, indicating that these variables have different effects on targets than on non-critical lures. The interaction between word frequency and semantic relatedness approached but did not reach statistical significance, $F(3, 213) = 2.51$, $MSE = 0.015$, $p = .06$. There was no evidence for a three-way interaction, $F(3, 213) = 1.53$, $MSE = 0.019$, $p > .10$. The two-way interactions were further broken down using separate ANOVAs for targets and non-critical lures along with sets of planned contrast comparisons.

Targets. For targets, a 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA revealed a significant main effect of semantic relatedness, $F(1, 71) = 4.64$, $MSE = 0.021$, $p = .035$, such that participants correctly identified studied words from semantically associated sets as old, $M = 0.80$, $SE = 0.011$, more often than words from

non-associative sets, $M = 0.77$, $SE = 0.010$. There was also a significant main effect of word frequency, $F(3, 213) = 9.42$, $MSE = 0.021$, $p < .001$, but there was no hint of an interaction, $F < 1$. Planned comparisons showed that the effect of word frequency on hit rates was driven exclusively by the LF condition. Participants correctly recognized LF targets at higher rates than MF targets, $t(71) = 3.49$, $SEM = 0.017$, $p < .001$, HF targets, , $t(71) = 4.41$, $SEM = 0.018$, $p < .001$, or VHF targets, $t(71) = 3.91$, $SEM = 0.020$, $p < .001$. However, there were no significant differences in participants' recognition of MF, HF, or VHF targets, all ts < 1.4 and all ps > .10.

Non-critical lures. For lures, a different pattern emerged. A 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA showed a large main effect of word frequency, $F(3, 213) = 51.06$, $MSE = 0.022$, $p < .001$, and a significant interaction between semantic relatedness and word frequency, $F(3, 213) = 3.50$, $MSE = 0.017$, $p = .016$, but no main effect of semantic relatedness, $F < 1$, on false alarms to non-critical lures. Because the interaction was significant, the effects of word frequency were analyzed separately for associative and non-associative lists.

For associative lists, LF non-critical lures were incorrectly judged to be old less often than were MF non-critical lures, $t(71) = -5.27$, $SEM = 0.018$, $p < .001$, and HF lists, $t(71) = -6.43$, $SEM = 0.020$, $p < .001$, but there was no difference in the false alarm rates for MF and HF lists, $t(71) = -1.46$, $SEM = 0.021$, $p > .10$. The largest change in false alarm rates was between HF and VHF lures, $t(71) = 4.62$, $SEM = 0.024$, $p < .001$. Thus, although there was an increase in false alarm rates for associative lists as word

frequency increased, the increase was in steps from LF to MF-HF and from MF-HF to VHF.

A similar step pattern was observed for non-associative lists, except it was in steps from LF to MF and MF to HF-VHF. For non-associative lists, the false alarm rate increased significantly from LF to MF, $t(71) = 5.20$, $SEM = 0.019$, $p < .001$, and from MF to HF, $t(71) = 2.84$, $SEM = 0.024$, $p = .006$. There was no difference in the false alarm rates for HF and VHF lists, $t(71) = 0.41$, $SEM = 0.027$, $p > .10$.

Old-new judgments: Semantically induced false recognition. The final set of analyses based on old-new judgments examines the effects of word frequency and prior study of associates on semantically induced false recognition of the critical lures to associative lists. A 2 (Prior Study: studied list vs. unstudied list) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA on false alarms to critical lures showed that there were main effects of prior study, $F(1, 71) = 11.27$, $MSE = 0.060$, $p = .001$, and word frequency, $F(3, 213) = 31.09$, $MSE = 0.053$, $p < .001$. These variables did not significantly interact, $F(3, 213) = 1.77$, $MSE = 0.066$, $p > .10$.

Effect of prior study and test on false recognition. Consistent with effects reported by Coane and McBride (2006) and Kimball et al. (2010) in which testing items from associative lists increased the false alarm rate to critical lures for those lists, participants falsely judged critical lures associated to unstudied but tested lists as old, $M = 0.31$, $SE = 0.017$, at a rate that was significantly higher than the false alarm rate for the list items themselves, $M = 0.25$, $SE = 0.13$, $t(71) = 3.69$, $SEM = 0.016$, $p < .001$. False alarms rates to critical lures increased even further, $M = 0.38$, $SE = 0.018$, when the associative list had been studied, $t(71) = 3.36$, $SEM = 0.020$, $p = .001$.

Effects of word frequency on false recognition. Separate sets of planned comparisons were used to examine the effects of word frequency on false recognition of lures from studied and unstudied lists. (Even though the interaction between prior study and word frequency was not significant, because the patterns appear to be qualitatively different I chose to perform the analyses separately, as originally planned). For critical lures to studied lists, planned comparisons showed that false alarms increased linearly as word frequency increased, $t(71) = 7.43$, $SEM = 0.127$, $p < .001$. The difference in false alarms to critical lures from LF and MF lists was not significant, $t(71) = 1.24$, $SEM = 0.037$, $p > .10$, but false alarms were higher for HF lists relative to MF lists, $t(71) = 2.11$, $SEM = 0.044$, $p = .038$, and higher still for VHF lists relative to HF lists, $t(71) = 3.37$, $SEM = 0.043$, $p = .001$. By contrast, the effect of word frequency on false alarms to critical lures for unstudied lists was driven exclusively by the VHF condition. Participants falsely endorsed critical lures associated to unstudied VHF lists at higher rates than those for unstudied HF lists, $t(71) = 3.86$, $SEM = 0.048$, $p < .001$, unstudied MF lists, , $t(71) = 4.15$, $SEM = 0.045$, $p < .001$, or unstudied LF lists, , $t(71) = 4.86$, $SEM = 0.042$, $p < .001$. There were no significant differences in false alarm rates for critical lures from unstudied LF, MF, or HF lists, all ts < 0.4 and all ps > .10.

ROCs. I next turn to analyses that take advantage of the full range of data available from participants' confidence ratings. As discussed in Chapter 2, ROCs can provide a more complete picture of participants' performance by showing the relationship between confidence and accuracy. Because the pattern of results from the analyses of old-new responses indicated that semantic relatedness and word frequency interact, I created separate ROCs for each level of word frequency for non-associative

lists (see Figure 7) and associative lists (see Figure 8). These ROCs were created from the means for each condition averaged across participants. Because of the relatively large number of within-subject conditions (8) and the relatively small number of stimuli per condition for each subject (6), I was not able to create ROCs for a large proportion of the individual participants. The individual ROCs for participants for whom an ROC could be constructed appear to be qualitatively similar to the ROCs created from the overall means, so we can assume that the aggregated ROCs are representative of the individual subject ROCs (Yonelinas & Parks, 2007). Nevertheless, they should be interpreted with some caution (Wixted, 2002).

As can be seen from a visual examination of Figures 7 and 8, overall the ROCs from this experiment appear consistent with the standard shapes of ROCs from previous item recognition studies (e.g., Yonelinas, 1994). The probability space ROCs are clearly non-linear and asymmetric, and the *z*-ROCs all appear to be fairly linear with slopes less than unity. There are also clear differences in the shapes of the ROCs as a function of word frequency, but it is not as clear from a visual inspection whether there are any systematic effects of semantic relatedness on the ROCs.

In order to examine such effects in more detail, I conducted a linear regression on the *z*-transformed cumulative hit and false alarm rates to determine the parameters of the *z*-ROCs and test for differences in these parameters as functions of semantic relatedness and word frequency. A full model/reduced model comparison confirmed that there were significant effects of these variables on both the slope and the intercept, $F(14, 24) = 109.16$, $MSE = 0.001$, $p < .001$. The slopes and intercepts of the *z*-ROCs as a function of semantic relatedness and word frequency are shown in Figures 9 and 10,

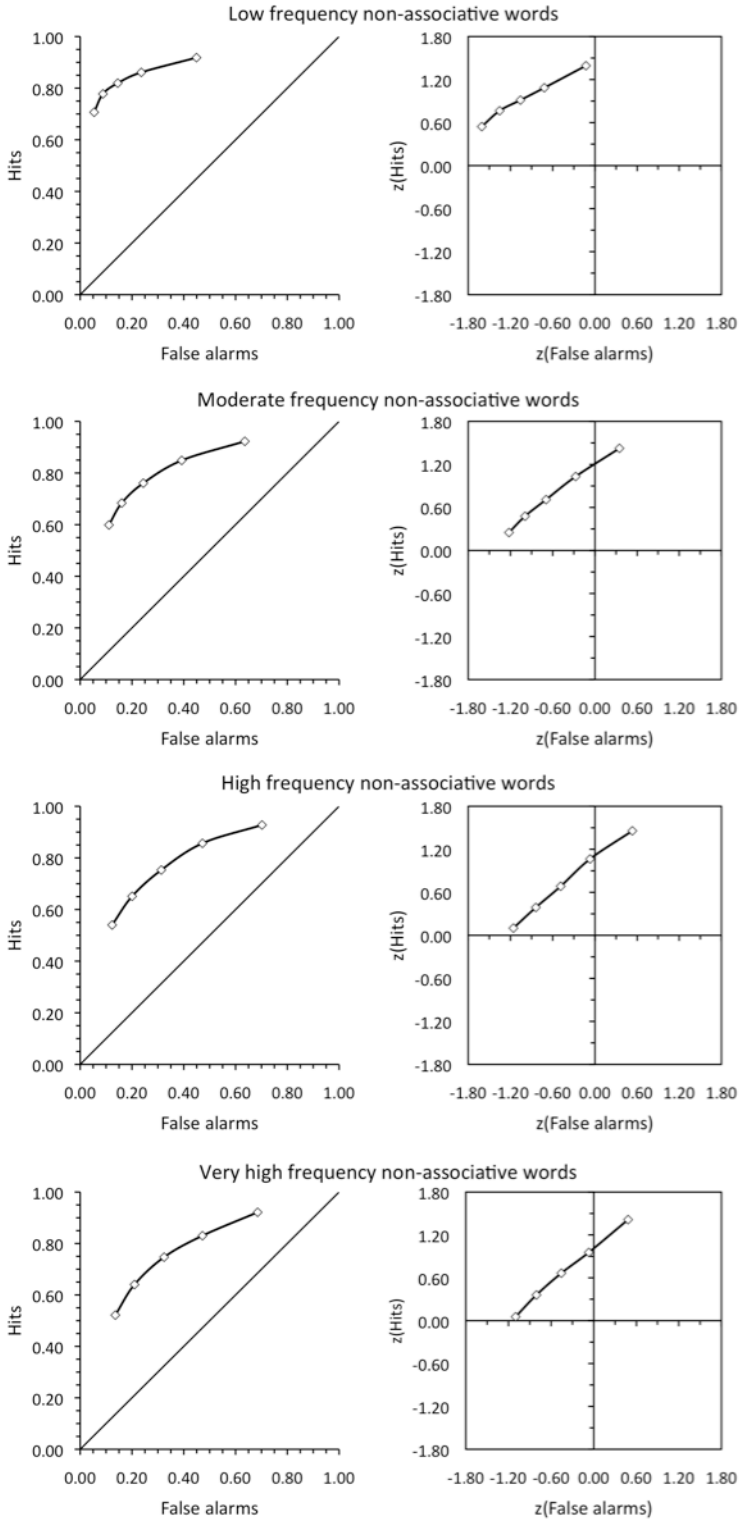

Figure 7. ROC and *z*-ROC curves for the words from non-associative word sets, by word frequency

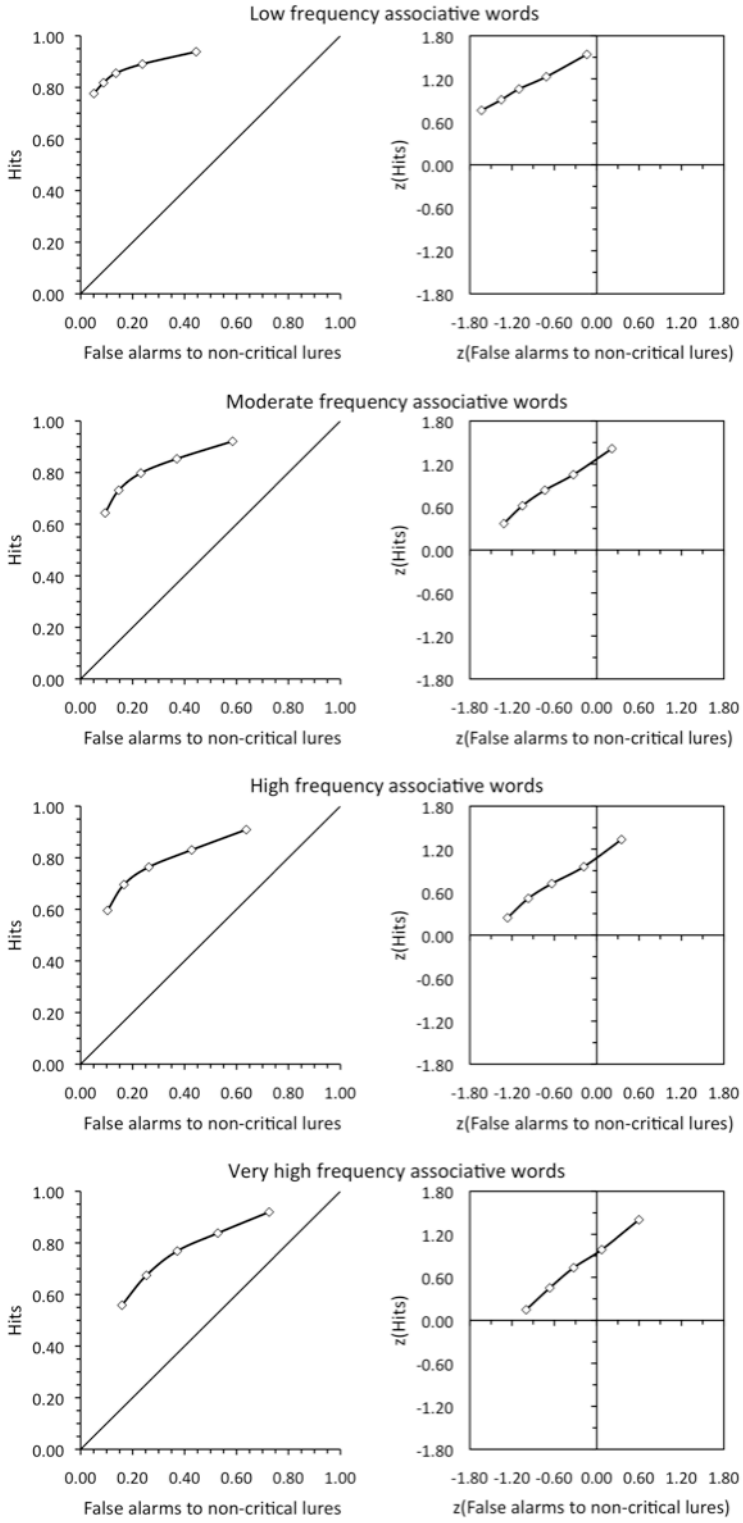

Figure 8. ROC and *z*-ROC curves for the words from associative word sets, by word frequency

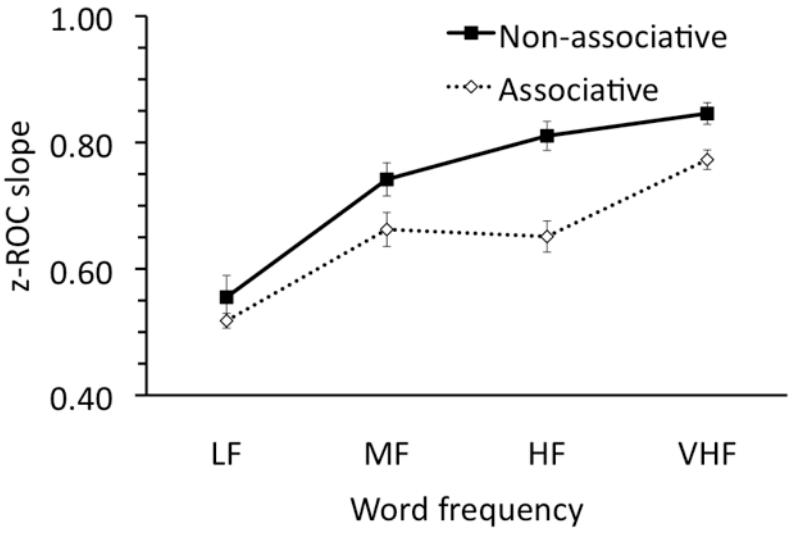

Figure 9. *z*-ROC slopes as a function of semantic relatedness and word frequency.

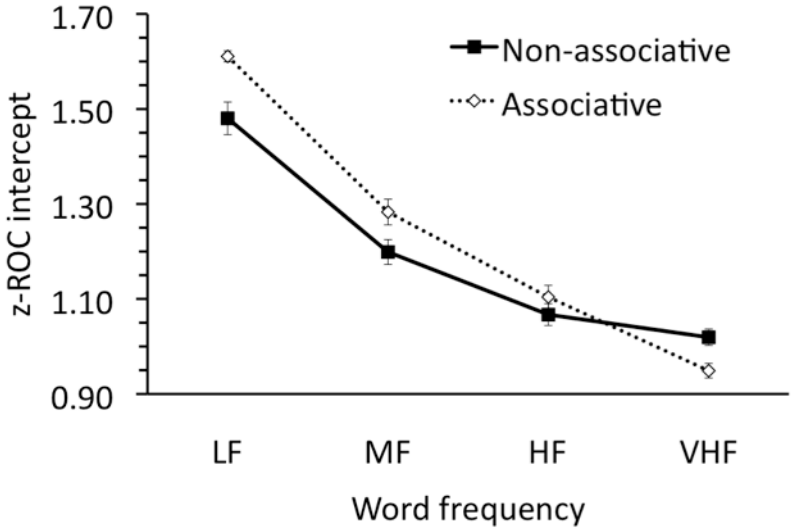

Figure 10. *z*-ROC intercepts as a function of semantic relatedness and word frequency.

respectively. As these figures clearly show, as word frequency increased the slopes of the *z*-ROCs also increased while the intercept decreased. But as with the old-new results, this is not a strictly linear relationship. The plots also clearly show that there were differences between the associative and non-associative lists that were moderated by word frequency. For low- and moderate frequencies, targets and lures from associative lists were more discriminable than were targets and lures from non-associative lists. But this gap closed as frequency increased, so that there was no difference in discriminability for HF words, and actually reversed for VHF words. This provides at least some support for the idea raised earlier that the pattern for false alarms to critical lures is due to low discriminability for VHF associative lists.

Fits to theoretical models. The ROCs can be interpreted in terms of either a signal detection model, such as the UVSD model proposed by Wixted (2007a), or in terms of the DPSD model (Yonelinas, 1994). In fitting the data with either model, it is necessary to make certain assumptions regarding the means and variances for the stimulus classes and the criteria settings for the experimental conditions. Because frequency and semantic relatedness were both manipulated within lists, it is reasonable to assume that the criteria were the same for all conditions. Therefore, in fitting the UVSD and DPSD models, I fit all 8 experimental conditions simultaneously and constrained the criteria to be the same for all conditions. To provide a base model for comparison, I first fit the 80 data points to an equal variance SDT model, using a standardized normal distribution for lures ($M = 0.0$, $SD = 1.0$) and a single distribution for targets. This model only used 1 degree of freedom but did not fit the data particularly well (SSE = 0.5200, RMSEA = 0.081).

UVSD interpretation. I next fit the UVSD model (Wixted, 2007a) to the cumulative hit and false alarm rates for each level of confidence, allowing both the means and variances of the target and lure distributions to vary for each experimental condition. Overall, the UVSD model fit the data extremely well (*SSE* = 0.0032, RMSEA = 0.008). The estimated parameters for the means and variances for each experimental condition are shown in Table 3. As can be seen from the table, the UVSD model predicts that the lure distributions have similar variances (between 0.68 and 0.76) that do not vary substantially with either word frequency or semantic relatedness. By contrast, the target distribution variances show a substantial effect of word frequency and a smaller but consistent effect of semantic relatedness. As word frequency increases, the variance of the target distribution decreases, with the largest difference being between LF and MF words. This is generally true for both associative and non-associative lists, although the associative targets exhibit slightly more variability overall than non-associative targets. According to the UVSD model, these differences in variability among the target distributions are responsible for the effects of word frequency and semantic association on the *z*-ROC slope shown in Figure 9.

The UVSD model also shows effects of word frequency and semantic relatedness on the means of the distributions. Interestingly, the means of the distributions do not show mirror effect patterns for word frequency predicted by Bayesian likelihood models (see Figure 5), even though these patterns are clearly evident in the hit and false alarm rates. Instead, the UVSD model orders both the lure and target distributions such that LF > MF > HF ≥ VHF. Consistent with the

Table 3. Estimated UVSD parameters in Experiment 1 as a function of word frequency and semantic relatedness.

Parameter		LF	MF	HF	VHF
Associative lists					
Lures	Mean	0.45	0.17	0.07	-0.12
	SD	0.76	0.75	0.70	0.72
Targets	Mean	1.99	1.25	1.13	0.99
	SD	1.52	1.10	1.08	0.90
Non-associative lists					
Lures	Mean	0.43	0.10	-0.03	-0.03
	SD	0.76	0.70	0.68	0.73
Targets	Mean	1.68	1.08	0.92	0.90
	SD	1.46	1.00	0.81	0.84

recollection hypothesis, the WFME is generated by a combination of differences in means for lures and differences in variances for targets.

Dual-process interpretation. A dual-process interpretation considers the contribution of familiarity and recollection—two qualitatively different types of memory and decision processes—when examining the shapes of the ROCs. To estimate the relative contributions of these processes according to the DPSD model (Yonelinas, 1994), I fit the DPSD model to the cumulative hit and false alarm rates from all 8 experimental conditions, again constraining the model to use the same criteria for all conditions but allowing for different values for the means of lure and target distributions and for the probabilities of recollection for each condition. Following the standard practice for the DPSD model, the standard deviations were set to 1.0 for all distributions (Yonelinas, 1994, 1997). The estimated familiarity and recollection

parameters for each experimental condition are shown in Table 4. The DPSD did not fit the data as well as the UVSD model did, but it still provided a reasonable fit (*SSE* = 0.0064, RMSEA = 0.011). However, the DPSD model shows a very different picture than does the UVSD model.

First, the DPSD models shows a clear mirror effect pattern (see Figure 5) in the estimates for familiarity, with the lure familiarities being ordered LF > MF > HF ≥ VHF (like the means in the UVSD model) but the target familiarities being in the reverse order for three of the word frequencies, LF < MF < VHF. The familiarity estimates for the HF condition were somewhat odd, with HF targets being less familiar than either MF or VHF targets for associative lists but more familiar than MF and VHF targets for non-associative lists. If familiarity is thought of as representing global memory strength

Table 4. Estimated DPSD parameters in Experiment 1 as a function of word frequency and semantic relatedness.

Parameter		LF	MF	HF	VHF
Associative lists					
Lures	Familiarity	0.30	-0.02	-0.08	-0.37
	Recollection	0.11	0.06	0.00	0.01
Targets	Familiarity	0.54	0.87	0.79	1.06
	Recollection	0.73	0.50	0.46	0.30
Non-associative lists					
Lures	Familiarity	0.30	-0.03	-0.21	-0.24
	Recollection	0.09	0.00	0.00	0.02
Targets	Familiarity	0.59	0.91	1.17	1.09
	Recollection	0.65	0.41	0.21	0.23

as proposed by Yonelinas (1994), then this pattern clearly violates the basic tenets of the SDT portion of DPSD theory (Glanzer & Adams, 1985). On the other hand, the ordering of familiarity distributions is roughly consistent with predictions from Bayesian likelihood models, but in order to account for the HF condition, these models would have to specify the similarities and differences between LF, MF, HF, and VHF words in more detail than they currently do (see, e.g., Shiffrin & Steyvers, 1997).

Second, the DPSD model also shows evidence for differences in recollection as a function of word frequency and semantic relatedness. There is an overall trend such that the recollectability of targets decreases as word frequency increases. There was also evidence of phantom recollection for LF and MF non-critical lures from associative lists as well as LF lures from non-associative lists. Together, these results provide converging evidence that recollection is moderated by word frequency, as assumed by the recollection hypothesis. There were also differences in recollection between associative and non-associative lists, with semantic relatedness increasing the probability of recollection by 7-9 points for LF, MF, and VHF words and by 26 points for the anomalous HF words.

Summary

Experiment 1 was designed to test differential predictions from the Bayesian likelihood hypothesis and the recollection hypothesis. The Bayesian likelihood hypothesis predicted mirror effects for word frequency such that false alarm rates would be ordered LF > MF > HF > VHF while hit rates would be in the reverse order, LF < MF < HF < VHF. The recollection hypothesis predicted that a WFME would only obtain if there were a difference in recollectability. Both hypotheses predicted a mirror

effect for semantic relatedness that did not obtain. The results do not clearly support one hypothesis over the other, but they are more consistent with the recollection hypothesis than the Bayesian likelihood hypothesis for two reasons.

First, although there were WFMEs for some comparisons, the WFME did not universally obtain as predicted by the Bayesian likelihood hypothesis even though there were word frequency effects. By manipulating word frequency over 4 levels instead of 2, I was able to observe differential effects of word frequency on hit rates and false alarm rates. Overall, false alarm rates to non-critical lures increased with word frequency, consistent with the recollection hypothesis' assumption that false alarms are due to differences in familiarity. Hit rates exhibited a different pattern in which there was a drop in correct recognition from LF to MF, but no difference between MF, HF, and VHF. This step pattern is not consistent with the Bayesian likelihood hypothesis but is consistent with the recollection hypothesis if one assumes that a) LF words are more recollectable than are MF, HF, and VHF words and b) there are no differences in recollectability between MF, HF, and VHF words.

Second, evidence from analyses of ROCs suggests that there are indeed differences in recollectability of words as a function of normative frequency that match this pattern. Estimates from the UVSD model showed that—assuming a single SDT decision process—the variances of the memory strength distributions for targets are dependent on word frequency such that lower frequency words have greater variability than higher frequency words. If the target distribution is the sum of the pre-experimental (lure) familiarity distribution, a study-induced familiarity distribution, and a recollective strength distribution, all with different means and variances, as suggested

by Wixted and Stretch (2004), these results imply that the amount of recollective strength that is added to an item during study is a function of the item's normative frequency (e.g., Gillund & Shiffrin, 1984; Glanzer et al., 1993; Shiffrin & Steyvers, 1997) and, further, that the variance of this strength is also a function of word frequency. Estimates from the DPSD model also imply that recollection is a function of word frequency and that LF words are more recollectable than are MF, HF, and VHF words.

Nevertheless, Experiment 1 was not able to provide conclusive evidence that the WFME is due to differences in the use of recollection between word sets of different frequencies. This was partly because the use of familiarity and recollection had to be estimated from aggregated ROCs. Therefore, in Experiment 2, I used a more direct measure of familiarity and recollection that could be analyzed at the individual level as well as the aggregate level.

Chapter 5 – Experiment 2

Experiment 2 was designed to investigate the effects of word frequency and semantic relatedness using the remember-know procedure. This procedure allows for a more direct measurement of familiarity and recollection processes than the estimation from ROC curves that was used in Experiment 1. Thus, Experiment 2 used the same design and procedures as Experiment 1, except that participants were asked to make 2-stage remember/familiar judgments instead of old-new confidence judgments.

Method

Participants. Participants were 71 undergraduate students enrolled in psychology courses at the University of Oklahoma who participated for partial course credit. All participants spoke and read English fluently. None of the individuals from Experiment 1 participated in Experiment 2.

Materials, design, and procedure. Experiment 2 used the same materials and design as were used in Experiment 1. The procedures were also identical with the exception that instead of using confidence judgments, a modified 2-stage remember-know procedure was used. For each test item, participants were asked to first make a binary old-new judgment by pressing the "O" or "W" keys, respectively. Following this judgment, they then indicated whether they had made the old-new judgment based on a sense of familiarity or on the ability to remember specific details from the study episode by pressing the "F" or "R" keys, respectively. As in Experiment 1, participants practiced making these judgments as part of the instructional phase before they began the experiment proper.

Results and Discussion

Old-new judgments: Effects of word frequency. Overall, Experiment 2 replicated the pattern of results from Experiment 1, with one key exception noted below (see Table 3 for means). A 3 (Item Type: target, lure, critical lure) x 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA revealed significant effects of item type, $F(2, 140) = 446.13$, $MSE = 0.065$, $p < .001$, and word frequency, $F(3, 210) = 16.56$, $MSE = 0.023$, $p < .001$, but these main effects were again qualified by a significant interaction, $F(6, 420) = 45.17$, $MSE = 0.013$, $p < .001$. As in Experiment 1, a set of planned contrasts revealed that the interaction was due to a pattern consistent with the presence of an overall word frequency mirror effect, with hit rates decreasing linearly as normative frequency increased, $t(70) = -8.10$, $SEM = 0.058$, $p < .001$, while false alarms to non-critical lures and critical lures both increased, $t(70) = 10.13$, $SEM = 0.060$, $p < .001$, and $t(70) = 7.13$, $SEM = 0.104$, $p < .001$, respectively.

However, unlike in Experiment 1, the relationship between word frequency and old judgments for targets was strictly linear. There was no evidence for a reliable quadratic trend for hit rates, $t(70) = 1.27$, $SEM = 0.021$, $p > .10$. LF targets were correctly judged to be old more often than MF targets, $t(70) = 4.60$, $SEM = 0.015$, $p < .001$; MF targets elicited higher hit rates than HF targets, $t(70) = 2.36$, $SEM = 0.016$, $p = .021$; and HF targets, in turn, elicited higher hit rates than VHF targets, $t(70) = 2.58$, $SEM = 0.016$, $p = .012$.

Non-critical lures showed a complementary pattern that replicated the results from Experiment 1. As in Experiment 1, false alarms to non-critical lures increased

Table 5. Mean proportions of items endorsed as old for targets (hits), non-critical lures (false alarms), and critical lures (critical lure FAs) as a function of word frequency and semantic relatedness in Experiment 2.

	Word frequency									
	Low		Medium		High		Very High			
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Overall										
Hits	0.86	(0.017)	0.79	(0.021)	0.76	(0.021)	0.72	(0.022)	0.78	(0.011)
False alarms	0.10	(0.012)	0.17	(0.019)	0.23	(0.022)	0.28	(0.021)	0.19	(0.010)
Critical lure FAs	0.20	(0.025)	0.19	(0.025)	0.26	(0.027)	0.43	(0.032)	0.27	(0.015)
Associative lists										
Hits	0.88	(0.018)	0.79	(0.024)	0.77	(0.024)	0.74	(0.024)	0.80	(0.012)
False alarms	0.10	(0.014)	0.18	(0.021)	0.22	(0.024)	0.30	(0.024)	0.20	(0.011)
Non-associative lists										
Hits	0.85	(0.020)	0.80	(0.022)	0.74	(0.023)	0.69	(0.027)	0.77	(0.012)
False alarms	0.10	(0.015)	0.15	(0.020)	0.23	(0.025)	0.26	(0.024)	0.19	(0.011)
Critical lure FAs										
Studied lists	0.23	(0.030)	0.23	(0.033)	0.32	(0.033)	0.49	(0.036)	0.32	(0.018)
Unstudied lists	0.17	(0.031)	0.15	(0.026)	0.19	(0.032)	0.37	(0.039)	0.22	(0.017)

with word frequency, and there was no evidence of a non-linear trend in false alarm rates to non-critical lures, $t(70) = -0.43$, $SEM = 0.024$, $p > .10$. LF lures were incorrectly judged to be old less often than MF lures, $t(70) = -5.18$, $SEM = 0.013$, $p < .001$. MF lures elicited lower false alarm rates than HF lures, $t(70) = -3.55$, $SEM = 0.016$, $p < .001$; and HF lures, in turn, elicited even lower false alarms than VHF lures, $t(70) = 2.90$, $SEM = 0.020$, $p = .005$.

The results for critical lures to associative lists also replicated the pattern from Experiment 1. In addition to the significant linear trend, there was a significant non-linear (quadratic) trend, $t(70) = 4.90$, $SEM = 0.037$, $p < .001$, due primarily to the lack of a significant difference in the probabilities of judging critical lures from low and moderate frequency associative lists to be old, $t(70) = 0.54$, $SEM = 0.022$, $p > .10$. Critical lures for HF lists were falsely recognized more often than those from LF lists, $t(70) = 2.14$, $SEM = 0.026$, $p = .036$, or MF lists, $t(70) = 2.66$, $SEM = 0.026$, $p = .010$, and critical lures for VHF lists were falsely recognized at rates even higher than those for HF lists, $t(70) = 5.23$, $SEM = 0.032$, $p < .001$.

Old-new judgments: Effects of semantic relatedness and word frequency. To test for effects of semantic relatedness and possible interactions between semantic relatedness and word frequency when using the remember-know procedure, I again conducted separate analyses using just the targets and the non-critical lures (see Table 3 for means). A 2 (Item Type: target vs. non-critical lure) × 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA revealed significant main effects of item type, $F(1, 70) = 626.98$, $MSE = 0.157$, $p < .001$, and semantic relatedness, $F(1, 70) = 5.89$, $MSE = 0.020$,

p = .012, on participants' willingness to judge an item as old. Unlike in Experiment 1, the main effect of word frequency on old-new judgments was not statistically significant, $F < 1$, but there was a significant interaction between item type and word frequency, $F(3, 210) = 86.63$, $MSE = 0.016$, $p < .001$. There were no other significant interactions, all Fs < 1.25 and all ps > .10. Due to the interaction between item type and word frequency, separate ANOVAs along with sets of planned contrast comparisons were conducted for targets and non-critical lures.

Targets. The effects of semantic relatedness and word frequency on correct recognition of targets were almost identical to those reported in Experiment 1, with the exception of a linear, rather than a non-linear, effect of word frequency. A 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA for targets revealed a small but significant main effect of semantic relatedness, $F(1, 70) = 5.35$, $MSE = 0.018$, $p = .024$, such that participants correctly identified studied words from semantically associated sets as old, $M = 0.80$, $SE = 0.012$, more often than words from non-associative sets, $M = 0.77$, $SE = 0.012$. The main effect of word frequency was also significant, $F(3, 210) = 28.74$, $MSE = 0.019$, $p < .001$, but again there was no hint of a Semantic Relatedness × Word Frequency interaction, $F(3, 210) = 1.23$, $MSE = 0.015$, $p > .10$. As described above, the effect of word frequency on hit rates was strictly linear, with participants correctly recognizing more LF targets than MF targets, more MF targets than HF targets, and more HF targets than VHF targets.

Non-critical lures. The main effects of semantic relatedness and word frequency on false recognition of non-critical lures were also similar to those reported in

Experiment 1, but unlike in Experiment 1 there was no interaction. A 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA on non-critical lures showed only a main effect of word frequency, $F(3, 210) = 40.99$, $MSE = 0.021$, $p < .001$. Semantic relatedness did not significantly effect false alarm rates to non-critical lures, $F(1, 70) = 2.28$, $MSE = 0.013$, $p > .10$, nor was there a significant interaction between semantic relatedness and word frequency, $F(3, 210) = 1.16$, $MSE = 0.014$, $p > .10$. Even though the interaction was not significant, the effects of word frequency were analyzed separately for associative and non-associative lists to facilitate comparisons with Experiment 1.

As in Experiment 1, false alarm rates for associative lists increased in steps from LF to MF-HF and from MF-HF to VHF as word frequency increased. LF non-critical lures were incorrectly judged to be old less often than were MF non-critical lures, $t(70) = -4.37$, $SEM = 0.018$, $p < .001$, HF lures, $t(70) = -6.19$, $SEM = 0.019$, $p < .001$, and VHF lures, $t(70) = -8.25$, $SEM = 0.019$, $p < .001$. The difference in the false alarm rates for MF and HF lists did not reach the level of statistical significance, $t(70) = -1.74$, $SEM = 0.023$, $p = .089$, but there were more false alarms to VHF lures than HF lures, $t(70) = 2.83$, $SEM = 0.028$, $p = .006$.

The pattern for non-associative lists was also the same as observed in Experiment 1, with false alarms to non-critical lures increasing as a function of word frequency in steps from LF to MF and MF to HF-VHF. The false alarm rate for non-associative lists increased significantly from LF to MF, $t(70) = 2.98$, $SEM = 0.018$, $p = .004$, and from MF to HF, $t(70) = 3.86$, $SEM = 0.020$, $p < .001$. There was no

significant difference in false alarm rates for HF and VHF lures, $t(70) = -1.50$, $SEM = 0.023$, $p > .10$.

Old-new judgments: Semantically induced false recognition. Significant effects of word frequency and prior study of associates on semantically induced false recognition of the critical lures to associative lists were found in Experiment 1, and this finding was replicated in Experiment 2. A 2 (prior study of semantic associates: studied list vs. unstudied list) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA on false alarms to critical lures showed that there were main effects of prior study, $F(1, 70) = 36.52$, $MSE = 0.037$, $p < .001$, and word frequency, $F(3, 210) = 29.37$, $MSE = 0.058$, $p < .001$. These variables did not significantly interact, $F > 1$.

Effects of prior study and test on false recognition. Although semantically induced false recognition was observed in Experiment 2, there was no test-induced increase in false recognition. Participants falsely identified critical lures to associative lists as old more often when the associative list had been studied, $M = 0.32$, $SE = 0.018$, than when it had only been tested, $M = 0.22$, $SE = 0.17$, $t(70) = 6.04$, $SEM = 0.016$, $p < .001$. However, unlike in Experiment 1, there was no significant difference in false alarms to critical lures from unstudied associative lists and false alarms to the list items themselves, $M = 0.20$, $SE = 0.011$, $t(70) = 1.31$, $SEM = 0.015$, $p > .10$.

Effects of word frequency on false recognition. To further compare the results with Experiment 1, the effects of word frequency on false recognition of critical lures from studied and unstudied lists were examined separately. For critical lures to studied lists, planned comparisons showed that false alarms increased linearly as word frequency increased, $t(70) = 6.81$, $SEM = 0.127$, $p < .001$, but there was also a

substantial non-linear component, $t(70) = 2.78$, $SEM = 0.059$, $p = .007$. As in Experiment 1, there was no difference in false alarms to critical lures from LF and MF lists, $t(70) = 0$, but false alarms were higher for lures from HF lists than those from MF lists, $t(70) = 2.69$, $SEM = 0.035$, $p = .009$, and higher still for lures from VHF lists, $t(70) = 3.48$, $SEM = 0.047$, $p = .001$.

The pattern for critical lures associated to unstudied lists further replicated the results from Experiment 1, with the VHF condition driving the word frequency effect. Participants false identified critical lures associated to unstudied VHF lists at higher rates than those for unstudied HF lists, $t(70) = 4.92$, $SEM = 0.035$, $p < .001$, unstudied MF lists, , $t(70) = 5.53$, $SEM = 0.039$, $p < .001$, or unstudied LF lists, , $t(70) = 4.82$, $SEM = 0.040$, $p < .001$. There were no significant differences in false alarm rates for critical lures from unstudied LF, MF, or HF lists, all $|t\text{s}| < 1.30$ and all $p\text{s} > .10$.

Remember-familiar judgments. Overall, the results from old-new judgments in Experiment 2 closely replicated the findings from Experiment 1, although there were a few notable differences. I now turn to analyses of participants' phenomenological judgments for items that were identified as old. Remember-familiar judgments can be analyzed using the same methods as are used to examine old-new judgments. Because "familiar" and "remember" judgments are complementary in the same way that "old" and "new judgments are, only one set of judgments needs to be analyzed. I chose to analyze "remember" judgments because these judgments are generally indicative of the use of a recollection process, and it is the use of recollection that is of particular interest in explaining word frequency mirror effects. Table 6 shows the mean proportion of test

items endorsed as "old" for which participants indicated an ability to remember specific details rather than a sense of familiarity.

Effects of word frequency. A 3 (Item Type: target, lure, critical lure) x 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA on these judgments revealed a significant main effect of item type, $F(2, 137) = 95.26$, $MSE = 0.106$, $p < .001$, but no significant effect of word frequency, $F(3, 210) = 1.74$, $MSE = 0.049$, $p > .10$. These findings were qualified by a significant interaction, $F(6, 287) = 6.21$, $MSE = 0.060$, $p < .001$. A set of planned comparisons was conducted to examine the differences in the simple effects of word frequency on remember judgments for studied items, non-critical lures, and critical lures.

For studied words, remember judgments decreased linearly with word frequency, $t(70) = 8.41$, $SEM = 0.081$, $p < .001$. LF targets were judged to be old based on recollection more often than were MF targets, $t(70) = 3.04$, $SEM = 0.017$, $p = .003$. MF targets were "remembered" more often than HF targets, $t(70) = 4.09$, $SEM = 0.021$, $p < .001$, and HF targets were remembered more often than VHF targets, $t(70) = 3.40$, $SEM = 0.019$, $p = .001$. This pattern is almost identical to the pattern of hit rates (compare row 1 in Table 4 to row 1 in Table 3), and is consistent with predictions from the recollection hypothesis. Thus, one possible explanation for the effects of word frequency on overall hits is that there are systematic differences in the recollectability of words as a function of word frequency and that these differences drive the change in hit rates. Also as predicted by the recollection hypothesis, there were no effects of word frequency on remember judgments for non-critical lures, all $|t\text{s}| < 1$ and all $p\text{s} > .10$.

Table 6. Mean probabilities of items judged to be "old" that were given "remember" judgments in Experiment 2. Standard errors are in parentheses.

	Word frequency									
	Low		Medium		High		Very High			
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Overall										
Studied words	0.80	(0.026)	0.74	(0.026)	0.66	(0.030)	0.60	(0.032)	0.70	(0.015)
Non-critical lures	0.27	(0.052)	0.30	(0.047)	0.30	(0.047)	0.29	(0.039)	0.29	(0.023)
Critical lures	0.36	(0.061)	0.33	(0.062)	0.49	(0.057)	0.41	(0.047)	0.40	(0.028)
Associative Lists										
Studied words	0.83	(0.026)	0.78	(0.029)	0.71	(0.031)	0.58	(0.036)	0.73	(0.016)
Non-critical lures	0.40	(0.073)	0.29	(0.055)	0.29	(0.051)	0.31	(0.048)	0.32	(0.028)
Nonassociative Lists										
Studied words	0.77	(0.032)	0.71	(0.031)	0.61	(0.036)	0.60	(0.035)	0.67	(0.017)
Non-critical lures	0.18	(0.055)	0.31	(0.058)	0.31	(0.054)	0.30	(0.047)	0.28	(0.027)
Critical lures										
Studied lists	0.34	(0.074)	0.41	(0.078)	0.54	(0.063)	0.51	(0.057)	0.46	(0.033)
Unstudied lists	0.42	(0.087)	0.22	(0.077)	0.35	(0.082)	0.30	(0.059)	0.32	(0.037)

Interestingly, the pattern of remember judgments for critical lures to associative lists was opposite the pattern for studied items: Remember judgments for critical lures that were falsely identified as old increased with word frequency, $t(29) = 2.06$, $SEM = 0.233$, $p = .048$. There were no significant differences in remember judgments between critical lures to LF and MF lists, $t(31) = 0.97$, $SEM = 0.061$, $p > .10$, or between critical lures to HF and VHF lists, $t(47) = 0.55$, $SEM = 0.074$, $p > .10$. But there was a significant difference between LF and MF lists and HF and VHF lists, $t(29) = 2.37$, $SEM = 0.118$, $p = .025$. Because there were only 6 critical lures for each level of word frequency and the mean false alarm rate for these lures was less than 0.30, contrasts for critical lure remember judgments could not be calculated for all participants; these results should, therefore, be interpreted with caution. Nevertheless, these patterns are consistent with other studies that have shown that the critical lure can rise to conscious awareness during processing of the associates and then be remembered on a later test based on that experience (e.g., Seamon, Lee, Toner, Wheeler, Goodkind, & Birch, 2002).

Effects of semantic relatedness and word frequency. To test for effects of semantic relatedness and possible interactions between semantic relatedness and word frequency on the use of remember judgments, I conducted separate analyses using just the targets and non-critical lures that had been judged as old (see Table 3 for means). The results of a 2 (Item Type: target vs. non-critical lure) × 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA were similar to the results from the analysis of old-new judgments. There was a significant main effect of item type, $F(1, 69) = 152.32$, $MSE =$

0.187, $p < .001$, with participants giving remember judgments to studied words, M = 0.70, SE = 0.015, more often than to non-critical lures, M = 0.29, SE = 0.023. There was a significant main effect of semantic relatedness, $F(1, 70) = 12.42$, $MSE = 0.047$, $p < .001$, such that participants remembered words from associative lists more often than words from non-associative lists. And there was a significant main effect of word frequency on remember judgments, $F(3, 210) = 4.79$, $MSE = .059$, $p = .003$, that was qualified by an interaction with item type, $F(3, 157) = 7.49$, $MSE = .058$, $p < .001$. There were no other significant interactions, all Fs < 1.60 and all ps > .10. As with old-new judgments, separate ANOVAs along with sets of planned contrast comparisons were conducted for targets and non-critical lures.

Semantic relatedness and word frequency had somewhat different effects on remember judgments than on old-new judgments. Both the main effects and the interaction in a 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA for targets were significant. Participants identified targets as old based on recollection more often for associative lists than for non-associative lists, $F(1, 70) = 13.12$, $MSE = 0.030$, $p < .001$. There was a significant main effect of word frequency on these judgments, $F(3, 210) = 37.94$, $MSE = 0.031$, $p < .001$, and a significant Semantic Relatedness x Word Frequency interaction, $F(3, 209) = 2.94$, $MSE = 0.037$, $p = .034$. Table 4 clearly shows that remember judgments for targets decreased as a function of word frequency for both associative and non-associative lists. The interaction was due to the one exception to this pattern—there was no difference between remember judgments for HF and VHF targets for non-associative lists.

The effects of semantic relatedness and word frequency were also different for non-critical lures. A 2 (Semantic Relatedness: associative vs. non-associative word sets) × 4 (Word Frequency: LF, MF, HF, VHF) repeated measures ANOVA on non-critical lures that were identified as old showed no significant effects of these variables on remember judgments. There was a marginal main effect of semantic relatedness, $F(3, 61) = 3.34$, $MSE = 0.069$, $p = .073$, but no evidence for an effect of word frequency or an interaction, Fs < 1.

***Semantically induced false recognition*.** There were not enough observations to conduct a reliable inferential analyses of the effects of word frequency and prior study on remember judgments for critical items that had been judged as old. Nevertheless, the means are presented in Table 4 and discussed in qualitative terms. From the means, it appears that prior study of the semantic associates to a critical lure increases the likelihood of that item being falsely "remembered", particularly for lists composed of high- and very high-frequency associates. There also appears to be an interaction for LF words, but given the small number of observations, this could be an outlier. To gather enough data to investigate these effects in more depth, a study would have to be designed in which each subject was presented with a large number of associative lists for each level of word frequency. This might be possible for low and moderate frequencies; but, unfortunately, there simply are not enough high- and very high-frequency words in the English language (or any language) to do this for these frequencies.

Discussion

Experiment 2 was designed to investigate the role of recollection in the word frequency mirror effect and in effects of semantic relatedness on item recognition. Evidence from Experiment 1 pointed to recollection as the locus of the WFME, with word frequency effects on false alarm rates being driven by differences in familiarity and word frequency effects on hit rates being driven by recollection, as predicted by the recollection hypothesis. The results from Experiment 2 provide further support for this hypothesis.

Overall, a consistent WFME was observed in Experiment 2, with hit rates being ordered LF > MF > HF > VHF and false alarm rates having the opposite order, LF < MF < HF <VHF. This is different from the pattern observed in Experiment 1, but based on the recollection hypothesis this difference is to be expected because of differences in the experimental procedures. In Experiment 1, participants were asked to provide confidence ratings; this task is relatively neutral in that it does not push the participants to make their recognition decisions using any particular cognitive process. But in Experiment 2, participants were explicitly asked to make a remember-familiar judgment. This explicit requirement can cause participants to rely more on recollection than on familiarity (Rotello et al., 2005). Thus, the observance of a limited mirror effect in Experiment 1 and a full mirror effect in Experiment 2 likely reflects differences in the use of recollection between the experiments, consistent with the recollection hypothesis. These findings are a challenge for the Bayesian likelihood hypothesis, though, because the only way in which it could accommodate these results is to suppose that the difference in experimental procedures somehow caused

participants to use different estimates of prior probabilities for HF and VHF items but not for LF and MF items (see Glanzer et al., 2009).

Participants' remember-familiar judgments in Experiment 2 also provide more direct evidence for impact of recollection on the WFME. Consistent with predictions from the recollection hypothesis, participants' remember judgments for studied items showed the same ordering as the hit rates. To verify that this is not an artifact of averaging over participants, I calculated the within-subject correlation between hit rates and remember judgments across the 4 levels of word frequency. Discarding the 4 participants for whom this was not calculable, the mean correlation was significantly positive, $M = 0.23$, $SE = 0.08$, $t(66) = 2.98$, $p < 0.01$.

There is one area where the results from Experiments 1 and 2 seem to contradict the recollection hypothesis. According to Joordens and Hockley (2000), the recollection hypothesis assumes that false alarms are driven by familiarity. This assumption has two consequences. First, it implies that there should be no differences in the use of recollection—whether measured from ROC parameter estimates or from remember judgments to lures—across the levels of word frequency. Although this is a prediction of a null effect, the evidence from Experiment 2 seems to support it; there were no significant differences in the use of recollection for non-critical lures. However, estimates from the DPSD model showing phantom recollection for LF non-critical lures (even for non-associative lists) in Experiment 1 suggest that recollection may play a role in generating false alarms for lures, at least some of the time.

Second, the assumption that false alarms are driven by familiarity implies that the advantage in low false alarm rates for LF words relative to HF words is a

consequence of LF words having a lower preexperimental familiarity than HF words (Joordens & Hockley, 2000). This assumption that word frequency is positively related to familiarity is a common assumption in memory research (e.g., Clark, 1992; Glanzer & Adams, 1985, 1990; Glanzer et al., 1993; Reder et al., 2000) because it seems psychologically intuitive, but the parameter estimates from both the UVSD and DPSD models in Experiment 1 contradict this assumption. According to this assumption, the means of the strength distributions (UVSD) and the familiarity distributions (DPSD) for lures should have been ordered as an increasing function of frequency (LF < MF < HF < VHF), but the estimates from the models reflected a decreasing function instead (LF > MF > HF > VHF). Thus, either both the models are wrong, or contrary to long held assumptions, word frequency is *negatively* related to familiarity. Although this conclusion seems counterintuitive, there is some evidence from neural network models of memory that supports it (Monaco et al., 2007).

Chapter 6

Computational Models Of Item Recognition

The Rise and Demise of Strength-Based Global Memory Models

One of the hallmarks of memory research in the 1980's was the development and use of global memory models (GMMs) to explain the processes that underlie recall and recognition memory performance. These models included the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984), the theory of distributed memory (TODAM2; Murdock, 1993), and MINERVA 2 (Hintzman, 1988). While these models differed widely in their architectures, they shared a common set of assumptions about how human memory operates (for a comparative review, see Clark & Gronlund, 1996). The GMMs all assumed that items are encoded into memory during study, and each of them specified the encoding process in mathematical terms. They also assumed that at test, a set of available cues were combined in short-term memory and used as a single, joint probe of long-term memory.

The defining characteristic of GMMs was that they assumed that this probe was compared to all items in memory and that a match value was calculated for each of these comparisons (Clark & Gronlund, 1996). This match value was variously called strength (e.g., Raaijmakers & Shiffrin, 1981), familiarity (e.g., Gillund & Shiffrin, 1984), or activation (e.g., Hintzman, 1988). As they were implemented initially, GMMs all assumed that recognition decisions are made by comparing the index of familiarity to a criterion and responding "old" if the familiarity was above the criterion or "new" if below it, similar to the assumptions of classical SDT. However, not all of

the GMMs were limited to this single-process view of recognition memory. For example, Gillund and Shiffrin explicitly included recall processes in the theoretical SAM model of recognition, although they only implemented and tested the single-process familiarity version.

Although the GMMs were able to explain a great variety of empirical data from human participants, a number of empirical challenges to the assumptions of recognition GMMs arose in the late 1980's and early 1990's (for reviews, see Clark & Gronlund, 1996; and Shiffrin & Steyvers, 1997). Mirror effects were foremost among these challenges (Glanzer & Adams, 1990), but other effects that posed difficulties for GMMs included the list length effect and list strength effects (Ratcliff, Clark, & Shiffrin, 1990), obtaining *z*-ROC slopes less than unity (Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992), dissociations in patterns from item and associative recognition tasks, and effects of verbal and environmental context (e.g., Clark & Shiffrin, 1992; S. Smith, Glenberg, & Bjork, 1978). Although each of the GMMs could account for a subset of these effects, no single GMM was able to account for all of them. As a consequence, many researchers—including some of the developers of the aforementioned GMMs—now consider the entire class of strength-based GMMs to have been falsified as theories of recognition (see, e.g., Diana et al., 2006; Glanzer et al., 2009).

Bayesian Likelihood Recognition Models

Attention/likelihood theory. As an alternative to the strength-based GMMs, Glanzer et al. (1993) developed attention/likelihood theory (ALT). Like many other computational models that derive from Estes (1955) stimulus sampling theory, ALT

assumes that stimuli consist of features, some of which are probabilistically sampled and marked as having been encountered during any given processing opportunity. For any given experiment, all items are assumed to start with some small proportion of their features having been marked from pre-experimental encounters. This noise marking probability is assumed to be the same for all stimuli. However, the number of features that are sampled during the experiment is assumed to be different for stimuli in different conditions, depending on the amount of attention that the stimulus class affords. During a recognition test, a sample of the test item features is drawn and the number of marked features is assessed. Unlike strength-based GMMs, though, ALT does not assume that the recognition decision is made on the basis of a direct comparison of the features. Instead, an additional step wherein the likelihood of the test item being new or old is made, and then this likelihood as the input to an SDT decision process.

In ALT, likelihood is assessed by using the number of marked features along with the subject's metacognitive knowledge[1] of how many marked features new and old items should be expected to have as the basis for calculating a Bayesian log likelihood ratio (Glanzer et al., 1993; Hintzman, 1994). Assuming the subject's metacognitive

---

[1] There is some dispute as to whether the type of knowledge that ALT assumes subjects use to perform the likelihood assessment is metacognitive in nature. Glanzer et al. (1993, pg. 551) state that "[t]he subject is assumed to have background information: how much marking a new [item] and how much marking an old item such as the given test item would have." Hintzman (1994) points out that this implicitly assumes certain metacognitive judgments. In their reply, Kim and Glanzer (1994) deny that this is metacognition, stating that the term "imports a large amount of theoretical baggage that we have no interest in handling" (pg. 206) and that ALT "simply assumes that the subjects estimate *p(i, old)* on the basis of their experience in the experiment" (pg. 207). However, Kim and Glanzer do not offer any process other than metacognition that would allow subjects to make this estimation, and until such time as an alternative process is proposed, metacognition seems to be the only viable process for this estimation.

knowledge is at least approximately accurate, this calculation has the natural consequence of creating log likelihood distributions that, when placed on a common decision axis, array themselves in such an order as to produce mirror effects (see Figure 11). That is, the lure distribution for the stronger stimulus class (AN) has a lower mean likelihood ratio than does the lure distribution for the weaker stimulus class (BN), and target distribution for the stronger stimulus class (AO) has a higher mean likelihood ratio than does the target distribution for the weaker stimulus class (BO). Thus, assuming that the subject chooses a criterion that is not too far from the optimal point, ALT predicts mirror effects as a natural byproduct of the recognition decision process (Glanzer et al, 1993). In fact, any model that uses the Bayesian likelihood as the basis for making recognition decisions will predict mirror effects when there are differences in the prior probabilities and an optimal or near-optimal criterion is used (Glanzer et al., 2009).

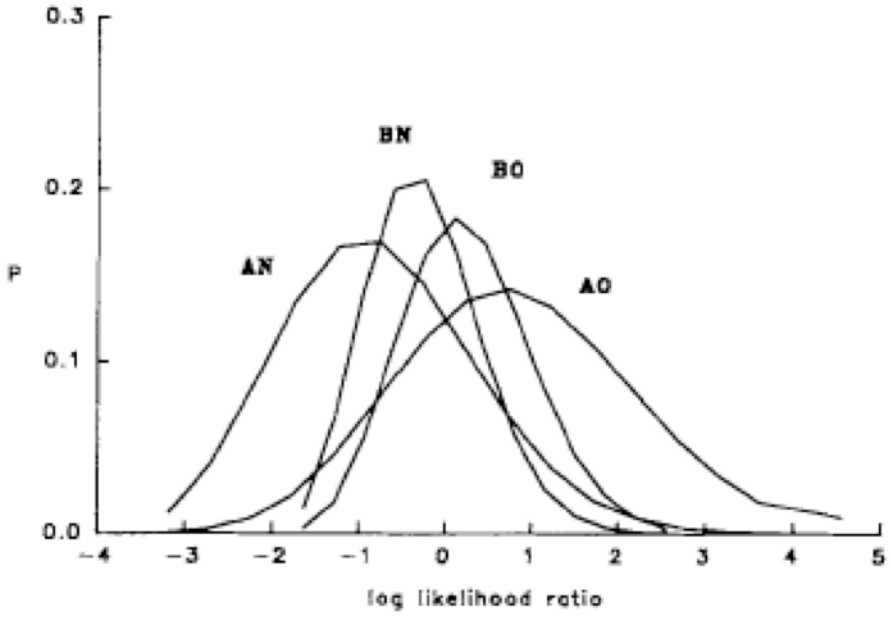

Figure 11. Distributions of Bayesian log-likelihood ratios in ALT (from Glanzer et al., 1993, Figure 3)

Although ALT has been able to account for mirror effects and a few other phenomena that appear when comparing stimuli or procedures that give rise to differences in recognition memory performance, the theory has been criticized on a number of fronts. For example, Hintzman (1994) argues that the theory makes a number of unstated assumptions, in particular relating to metamemory processes[1], and that other ways of rescaling strengths can produce the same effects. Maddox and Estes (1997) point out that in order to ensure that the mirror effect occurs properly, ALT requires assumptions about how the characteristics of the stimuli or encoding task relate to the number of features that are sampled. These assumptions are not always obvious *a priori,* and in some cases they do not make logical sense.

Retrieving effectively from memory (REM). REM was developed in reaction to the failure of the SAM model to account for mirror effects and ROC curves, and it overcomes the limitations of the SAM architecture by incorporating two significant changes (Shiffrin & Steyvers, 1997). Instead of using strength-based familiarity signals as the basis for recognition, REM first calculates the Bayesian likelihood ratio of the match between the copy cue and each item stored in memory and then sums these likelihood ratios to calculate the odds that the cue matches an item in memory. If the odds are above a threshold, then the copy cue is judged to be old; otherwise it is judged to be new. To support the calculation of likelihood ratios, the representation of items in memory was changed from the associative matrices used in SAM to sets of vectors of feature values. The use of Bayesian likelihood ratios ensures that mirror effects are always produced in REM when stimuli classes have different strengths in memory (Glanzer et al., 2009). For example, the REM model accounts for the WFME by

assuming that high-frequency words have more common features than do low-frequency words. This causes a difference in the match values for low- and high-frequency words that then translates to a mirror effect through the Bayesian likelihood transformation.

One of the strengths of REM is that it can account for a wide range of findings from single-item recognition paradigms as well as a limited number of recall phenomena. Recognition phenomena that REM has been shown to account for include the effects of variations in list length, list strength, and item strength on discriminability and ROC measures of item recognition (Shiffrin & Steyvers, 1997); the presence of mirror effects in word frequency experiments (Malmberg, Holden, & Shiffrin, 2004; Shiffrin & Steyvers, 1997); and various measures of remember-know performance (Malmberg, 2008). Additionally, REM has been used to account for accuracy and response time in cued recall (Diller, Nobel, & Shiffrin, 2001), the list strength effect in free recall (Malmberg & Shiffrin, 2005; Verde, 2009), and the use of a recall-to-reject strategy in recognition (Malmberg, 2008; Malmberg et al., 2004).

Despite the successes of the REM model, it has a number of critical limitations. First, the Bayesian decision process that forms the core of the REM model predicts that mirror effects are a consistent feature of recognition, despite evidence of boundary conditions for mirror effects (see Chapter 3). Second, REM cannot account for a number of key effects in associative recognition including episodic fan effects (Buchler, Light, & Reder, 2008) and the influence of semantic relations between the cue and target (Greene & Tussing, 2001). Third, without some kind of pre-experimental semantic memory such as the one used in the fSAM model, REM is unlikely to be able

to account for semantic phenomena such as intrusions in free recall or false alarms to semantically related items in recognition. Finally, although Shiffrin and colleagues have argued that REM is an extension of the SAM model (Klein, Shiffrin, & Criss, 2007; Raaijmakers, 2005), the ability of the REM model to provide a satisfactory account for many of the phenomena that SAM was able to explain has never been tested. For example, it is not known whether REM can explain serial position curves in free recall (Raaijmakers & Shiffrin, 1980; but see Lehman & Malmberg, 2009), the part-set cuing effect (Raaijmakers & Shiffrin, 1981), or interference in cued recall and recognition (Mensink & Raaijmakers. 1988).

Boundary conditions on mirror effects: A challenge for Bayesian likelihood models. The boundary conditions on mirror effects discussed in Chapter 3 and observed in Experiments 1 and 2 pose a challenge for Bayesian likelihood models such as ALT and REM. As discussed above, these models predict that under normal circumstances mirror effects will obtain for any set of stimuli where one stimulus class is stronger than the other. Glanzer et al. (2009) showed that Bayesian likelihood models predict exceptions to this rule when subjects exhibit a bias, but these predictions do not necessarily comport with the empirical evidence. For example, according to Glanzer et al., when subjects adopt a liberal criterion (i.e., one that is to the left of the optimal criterion), the strong stimuli should have a lower FAR than the weak stimuli but the HR for the two classes should be the same. Conversely, Glanzer et al. also showed that when subjects adopt a strong criterion (i.e., one that is to the right of the optimal criterion) Bayesian log likelihood models predict that strong stimuli should have a higher HR than the weak stimuli but the FAR for the two classes should be the same.

However, these are global criterion shifts that should affect all classes of items; therefore, they cannot explain results such as those in Experiment 1 or in Estes and Maddox (2002) where mirror effects obtain for some item classes but not for others. Additionally, these bias shifts do not allow for concordant patterns where hit rates and false alarm rates are both higher for one stimulus class than for another.

Model Development and Testing

As discussed above, no extant model of recognition memory can account for mirror effects and their boundary conditions. There are three potential solutions to this problem: 1) modify a strength-based model to account for the presence of mirror effects, 2) modify a Bayesian likelihood model to account for the absence of mirror effects, or 3) develop a new class of recognition models that uses something other than strength or Bayesian likelihood as the basis for recognition decisions. For this thesis, I chose the first approach and tried to modify a strength-based model that is built on the Search of Association Memory (SAM) framework (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981). I chose the SAM framework rather than a newer Bayesian likelihood model for the reasons outlined below.

Rationale for the modeling approach. Unlike most Bayesian likelihood models, the SAM model provides a clear theoretical link between the processes of recall and recognition (Gillund & Shiffrin, 1984). In addition to accounting for an impressive array of recognition phenomena, variants of the SAM model have been able to account for a large number of key phenomena in free and cued recall—many of which cannot be accounted for by other computational models—whereas the ability of Bayesian likelihood models to account for free recall data is unknown (but see Lehman &

Malmberg, 2009). The phenomena to which SAM has been applied include the effects of presentation rate and list length (Raaijmakers & Shiffrin, 1980), serial position curves and part-set cuing (Raaijmakers & Shiffrin, 1981), interference and forgetting (Mensink & Raaijmakers, 1988), list strength (Shiffrin, Ratcliff, & Clark, 1990), generation (Clark, 1995), temporal contiguity (Kahana, 1996), category clustering (Sirotin, Kimball, & Kahana, 2005), and the generation of false recall in the Deese-Roediger-McDermott (DRM) paradigm (Kimball et al., 2007). With the notable exception of the REM model, none of the Bayesian likelihood models include memory search processes to support recall. The REM model theoretically includes a memory search process based on the memory search algorithm used in SAM (Shiffrin & Steyvers, 1997, pp. 160-161), but until recently the implementation of recall processes in the model has been limited to supplementing recognition decisions (Malmberg et al., 2004; Xu & Malmberg, 2007; Malmberg, 2008). Lehman and Malmberg (2009) developed a recall version of REM and applied it to the directed forgetting paradigm, but the generality of this model is still unknown.

Second, despite the fact that the general SAM recognition model proposed by Gillund and Shiffrin (1984) included both a familiarity process and a recall-like process, the capability of a dual-process SAM model—or any dual-process strength-based recognition model—to account for mirror effects has never been tested. This may be due to the fact that when the current generation of computational models was developed in the 1980's and 1990's the zeitgeist was that a single signal-detection theory (SDT) process provided the most parsimonious explanation for recognition and that adding an additional process would needlessly complicate the models (Slotnick & Dodson, 2005).

However, as discussed in Chapters 2 and 3, there is now a growing consensus that a multi-process theory that integrates a familiarity mechanism and a recollection mechanism is necessary to account for the full range of recognition memory phenomena (for recent reviews, see Wixted, 2007a, and Yonelinas & Parks, 2007), and a number of researchers have suggested that a such a theory could account for mirror effects (e.g., Arndt & Reder, 2002; Balota, Burgess, Cortese, & Adams, 2002; Cary & Reder, 2003; Hirshman, Fisher, Henthorn, Arndt, & Passannante, 2002; Joordens & Hockley, 2000). Additionally, even the Bayesian likelihood models such as REM are now being supplemented with recall and/or recollection mechanisms in order to account for more subtle effects of variables that often produce mirror effects (Malmberg et al., 2004; Xu & Malmberg, 2007; Malmberg, 2008), thus negating the potential argument that a single-process Bayesian likelihood model is more parsimonious than a dual-process strength-based model.

Finally, with the addition of a pre-experimental semantic memory, the SAM framework could be used to examine the interaction of semantic and episodic influences on memory performance (Kimball et al., 2007), something no other extant recognition model can do. This is important when examining mirror effects because many of the variables that produce mirror effects, such as normative word-frequency, concreteness and imageability, and meaningfulness, are inherently semantic variables rather than episodic variables (Monaco et al., 2007). I also needed a model that could represent semantic relations in order to try to account for the effects of semantic association in the two experiments presented in Chapters 4 and 5. I next describe the fSAM model of

recall that I used as a starting point for the development of a new model of recognition memory.

The fSAM Model of Recall

The fSAM model is based on the Search of Associative Memory model (SAM; Raaijmakers & Shiffrin, 1981) that has been previously applied to a wide variety of episodic memory phenomena including free recall, cued recall, recognition, and forgetting. The fSAM model incorporates the basic episodic memory machinery of SAM described below and can account for basic episodic memory phenomena such as the serial position curve in free recall (Kimball, et al., 2007) and part-set cuing effects (T. Smith & Kimball, 2008). However, unlike other computational memory models, the fSAM model explicitly represents semantic associations in long-term memory and is able to simulate semantic memory effects by using those semantic associations to construct semantic context at encoding and to search memory at retrieval. In addition, to promote ecological validity, the semantic associations in fSAM can be based on behavioral word association norming data (e.g., Nelson et al., 2004; Steyvers et al., 2004), rather than just abstract representations of semantic information, as are used in other models such as MINERVA (Arndt & Hirshman, 1998) or REM (Malmberg et al., 2004).

The SAM model of episodic recall. The SAM model (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981) assumes the existence of two memory stores: short-term memory (STM) and long-term memory (LTM). Within STM, rehearsal processes are idealized in the form of a limited capacity buffer in which studied words become associated through a rehearsal process, as described below. LTM contains

values for the strengths of two types of associations: the associations formed at study between each list word and the list context, and the pairwise episodic associations formed among list words during study. The strengths of item-to-context and inter-item associations formed during study are stored in an episodic matrix. List context is conceptualized as the temporal and situational setting for a particular list. For the sake of simplicity, the basic SAM model assumes that all associations in LTM are episodically created in the course of rehearsal during study, so the strengths in the episodic matrix are set to zero prior to study (although these associative strengths are later reset to a residual value for pairs of words that are not rehearsed together during study).

According to Raaijmakers and Shiffrin (1981), during study of a list, SAM assumes that, as each list item is presented, it enters the STM buffer and is rehearsed along with other items occupying the buffer at any given time, thereby increasing the strengths of the items' episodic associations in LTM. In particular, rehearsal increases the strength of association between each item in the buffer and the list context, the strength of the association in LTM between any two items that simultaneously occupy the buffer, and the association of each item in the buffer to itself. The amount of time that each item spends in the STM buffer during study is determined by the presentation rate, the size of the buffer (the maximum number of items that can simultaneously occupy the buffer), and the rule for displacement of items from the buffer. Once the buffer is full, each new item displaces one of the items then occupying the buffer.

Raaijmakers and Shiffrin (1981) describe recall from LTM in SAM as a self-limiting process in which memory is probed with a cue set and searched for items that

are associated to the cue set. The search process includes at least two phases: First, an item is sampled, and then it may or may not be recovered, that is, identified as a particular word and output as a recalled item. The sampling rule in SAM is a probabilistic Luce choice rule in which the strength of association between the cue set and each item in memory is compared to the sum of the associations between the cue set and all items in memory. This sampling rule uses the relative strength of items in memory to implement retrieval competition, so that items that are more strongly related to the cue set are more likely to be sampled. If an item is sampled, then its absolute strength of association to the cue set is used to determine whether the item will be recovered.

The fSAM framework. Building on the SAM model, Kimball et al. (2007) incorporated explicit representation of pre-experimental semantic associations among words, along with new semantic mechanisms that operate at encoding and retrieval, into a new model that was developed to account for veridical and false recall following the study of lists of semantically associated words. Kimball et al. developed several versions of fSAM that differed in the ways that these semantic associations exert an influence on episodic recall. One critical way that the model versions differed is in whether they incorporated a semantic mechanism at encoding, at retrieval, or at both stages. In addition, the particular semantic mechanism used at encoding was one of three different versions, and there were also three different versions of the semantic retrieval mechanism. In the different versions of the semantic encoding mechanism, each word in the lexicon becomes associated to the list context in proportion to the word's strength of semantic association either to the most recently presented word alone

or to all of the studied words jointly occupying the rehearsal buffer at a given time; if the latter, the association strengths combine either additively or multiplicatively. In a similar way, at test, the probability of retrieving a word is in part a function of its strength of semantic association either to the last recalled word alone or to all of the most recently recalled words jointly; if the latter, the strengths combine either additively or multiplicatively.

By factorially combining the encoding and retrieval mechanisms, Kimball et al. (2007) generated 16 model versions comprising a 4 (semantic encoding mechanism: none, single-item, additive, multiplicative) × 4 (semantic retrieval mechanism: none, single-item, additive, multiplicative) design and compared the ability of these 16 model versions to fit empirical data from a free recall experiment that they had run with human participants. They used the best-performing model version from those simulations—the version that combined the multiplicative encoding and retrieval mechanisms—to simulate important false memory effects from the literature, including developmental patterns of false recall and the effects of varying association strengths between and within study lists.

The multiplicative fSAM model used by Kimball et al. (2007) added three key features to the SAM framework. First, it extended the memory representations to include pre-experimental semantic associations along with the contextual associations and episodic interitem associations that were used in previous SAM-based models. Second, it added a mechanism to allow these semantic associations to impact the encoding of words during study. Third, it extended the concept of compound cuing to allow multiple items instead of just one item to be used to jointly cue memory along

with the current context. Together, these features give the fSAM model the ability to account for semantic effects in both veridical and false recall.

Kimball and Smith (2007) have further applied the fSAM model to other core false recall effects, including the effects of repeated study, repeated testing, repeated study-test trials, presentation rate, levels of processing, critical word presentation during study, and blocked versus random presentation of multiple lists. We have also applied the fSAM model to studies of part-set cuing and false memory (T. Smith & Kimball, 2008). However, the current fSAM model is still limited to explaining free recall phenomena.

Extending the fSAM Model to Recognition

Because the current version of the fSAM model lacks the mechanisms needed to implement decision processes that are vital to simulating recognition phenomena, the first step in extending the model to recognition was to develop a set of potential mechanisms to extend the fSAM model to recognition, as described below. For this thesis, I implemented each of these mechanisms and systematically tested the ability of the resulting 36 model variants to a) generate appropriately shaped ROCs in a simulation of a basic episodic recognition task and b) generate a word frequency mirror effect in a simulation of Glanzer and Adams (1990, Experiment 2), a seminal WFME experiment. I had planned to use the most successful models to try to explain the effects of semantic association and word frequency observed in Experiments 1 and 2 (see Chapters 4 and 5, respectively), but as it turned out none of the model variants was able to generate basic word frequency mirror effects. I explain these results in more detail in Chapter 7.

Recognition memory within the SAM framework. Gillund and Shiffrin (1984) proposed a general recognition model within the SAM framework that included both familiarity and recollection processes, and implemented a simple, single-process computer model to simulate a number of episodic recognition memory effects. In the Gillund and Shiffrin simulation model, recognition decisions were based on familiarity, much as in classic SDT models. Familiarity of a test probe was calculated by using the test item and context as cues, calculating for each item in LTM a product of that item's contextual association strength and its strength of association to the test item, then summing those products across all items in the lexicon to determine the familiarity of the test item. This familiarity value was the same as the denominator in the sampling rule used by Raaijmakers and Shiffrin (1981) in the recall process, thus uniting recognition and recall theoretically. If the calculated familiarity was above a pre-set criterion, the item was accepted as old. Otherwise it was rejected as being new.

Using this simple single-process model, Gillund and Shiffrin (1984) were able to qualitatively simulate a number of key findings from the recognition literature, including effects of presentation speed and retention interval, list-length effects for lists of unrelated words, and the mirror effect for word-frequency. However, in order to account for word-frequency effects, Gillund and Shiffrin assumed that participants use different criterion for high-frequency words than for low-frequency words, an assumption that no longer seems tenable in light of evidence that participants are reluctant to use different criteria even when given good reason to do so (Stretch & Wixted, 1998a).

Surprisingly, little attention has been given to the more general recognition theory laid out by Gillund and Shiffrin (1984) or the fact that they acknowledged the inherent limitations of their single-process simulation model as well as the arbitrariness of the method they used to calculate familiarity (see also Shiffrin & Steyvers, 1997). Because implementations of the SAM recognition model have generally assumed that recognition relies solely on a familiarity process (e.g., Clark & Shiffrin, 1992; Gillund & Shiffrin, 1984; Mensink & Raaijmakers, 1988; Shiffrin, Huber, & Marinelli, 1995) it is commonly thought that this assumption is integral to the SAM theory. However, a closer reading of the literature reveals that familiarity and recollection both play a role in recognition in the SAM framework, and there is a potential role for other decision processes as well. As Gillund and Shiffrin (pg. 56) put it, "the question is really not whether search processes occur, put the degree to which such processes occur…. [T]he underlying logic of the SAM model requires that the subject be able to utilize search if he or she so chooses."

Thus, there are at least two types of potential modifications that could be made to the traditional SAM-based approach to modeling recognition: a) use a different formulation to calculate familiarity and/or b) incorporate a recollection process into the model. To date, no studies of alternative ways to calculate familiarity within the SAM framework have been published, although Shiffrin et al. (1995) did add a differentiation mechanism that impacted the calculation of familiarity in response to lures. In the only published study that has attempted to incorporate recollection processes into the model, Ratcliff, Van Zandt, and McKoon (1995) showed that SAM could account for data from the process dissociation procedure when recall processes were allowed to contribute to

recognition memory decisions. Despite this success, no studies to date have implemented a full dual-process version of SAM. Thus, the ability of SAM-based models to account for mirror effects has never been fully tested.

Chapter 7

Simulations

My first step toward creating a strength-based model to account for mirror effects was to find a set of familiarity and recollection mechanisms that could be placed in the fSAM framework to generate ROCs that approximate the shape of human-generated ROCs in item recognition. Generating correctly shaped ROCs is a logical prerequisite for generating mirror effects because the same cognitive mechanisms are thought to underlie the two.

As discussed in Chapter 2, a signal detection model can only generate item recognition ROCs that look like those obtained from humans if the model assumes that the variance of the target distribution is larger than the variance of the lure distribution (Wixted, 2007). However, global memory models—and the SAM model in particular—have historically had difficulties generating asymmetric ROCs with -normalized slopes less than one (Ratcliff et al., 1992). Strength-based GMMs can produce distributions with equal variances (Murdock, 1993) or unequal variances where the variance of the lure distribution is larger than the variance of the target distribution (e.g., Gillund & Shiffrin, 1984), but in their current form they do not seem to be able to produce a target distribution that has a greater variance than the lure distribution. Wixted (2007a) has suggested that this could be accomplished by adding the strengths from a continuous-strength recollection process to familiarity. This type of operation conceivably could produce mirror effects as well. In particular, if the ordering of the recollection distributions is opposite that of the familiarity distributions—i.e., the more

familiar stimuli are less recollectable while the less familiar stimuli are more recollectable—the summed distributions could order themselves as shown in Figure 5, thereby producing a mirror effect when subjected to a signal detection process (cf., Reder et al., 2000).

Of course, asymmetrical ROCs with a normalized slope less than unity can also be created by supplementing familiarity with a discrete recollection process (Yonelinas, 1994; Yonelinas & Parks, 2007). Again, if one assumes that the more familiar stimuli are less recollectable while the less familiar stimuli are more recollectable, this combination of decision processes could produce a mirror effect. As Joordens and Hockley (2000) explain, decisions for lures necessarily are based on familiarity, so that the stimuli with lower familiarity will have lower false alarm rates. But decisions for targets can often be made using recollection. Because the less familiar stimuli are more recollectable, this will cause them to have higher hit rates than the more familiar stimuli.

Thus, a SAM model that can generate properly shaped ROCs—whether it uses a pure signal detection decision mechanism with unequal variances or a combination of recollection and familiarity decision mechanisms—might also be able to generate mirror effects using this same mechanism. Conversely, a model that cannot generate properly-shaped ROCs is unlikely to be able to generate the more complex patterns associated with mirror effects. In the next section, I describe a number of ways in which familiarity and recollection could be assessed within the fSAM framework and the results of tests of model variants that implement these mechanisms.

Modifications to the fSAM Model

The first step in extending the fSAM recall model to account for recognition was to add a mechanism to calculate and assess the familiarity, or strength, of items in memory. I tested four different mechanisms to calculate familiarity. Two of these mechanisms were based on summing the strengths of association between the memory probe and each item in memory over the entire contents of LTM; these are *global* familiarity mechanisms. The second two calculated familiarity based on the strength of the memory probe itself; these are *local* familiarity mechanisms. For each of these types of mechanisms, the strengths of the contextual, episodic, and semantic components of memory were combined either additively or multiplicatively. These mechanisms are described in more detail below.

Familiarity. The *global product* familiarity mechanism is an extension of the mechanism used by Gillund and Shiffrin (1984) in their episodic SAM model. In this model, familiarity is the denominator from the free recall sampling rule, wherein the strength of each item in memory is calculated as the weighted product of the three memory components as shown in Equation 1:

$$f(j) = \sum_{k}^{k \in N} \left[S(k,context)^{W_c} S_e(k,j)^{W_e} S_s(k,j)^{W_s} \right] , \qquad (1)$$

where *f(j)* represents the familiarity of the test probe *j*, $S(k, context)$ represents the strength of the association of item *k* to context; $S_e(k, j)$ represents the strength of the episodic association between the test probe *j* and item *k*; $S_s(k, j)$ represents the strength of the semantic association between the test probe and item *k*; W_c, W_e, and W_s are the retrieval weight parameters for weighting of item-to-context associations, inter-item

episodic associations, and inter-item semantic associations, respectively; and *N* is the set of all items in long-term memory. A slight modification of this rule yields the *global sum* familiarity mechanism:

$$f(j) = \sum_{k}^{k \in N} \left[W_c S(k,context) + W_e S_e(k,j) + W_s S_s(k,j) \right] \quad (2)$$

Although these two mechanisms assess the strength of each item in memory using a different rule, they both calculate familiarity by adding these strengths over all items in memory.

Although SAM has traditionally been identified as a global memory model, it is also possible to use a local strength to calculate familiarity. The *local product* familiarity mechanism uses the memory probe's self-association strength, combining the memory components multiplicatively:

$$f(j) = S(j,context)^{W_c} S_e(j,j)^{W_e} S_s(j,j)^{W_s} \quad (3)$$

If the memory components are combined additively instead, a *local sum* familiarity rule can be created:

$$f(j) = W_c S(j,context) + W_e S_e(j,j) + W_s S_s(j,j) \quad (4)$$

Each of the above formulations is theoretically plausible according to various conceptualizations of familiarity, and except for historical consistency with previous SAM-type models there is no apparent *a priori* reason to prefer any one to the others.

Recollection. The next step in creating a recognition model was to incorporate a recollection processes into fSAM. As with familiarity, there are a number of possible ways in which recollection processes can be implemented, and the manner in which the familiarity and recollection processes will be combined has to be considered, as well. I

tested a total of 8 different mechanisms for calculating recollection and combining the recollection and familiarity processes.

Recall-based memory search. The most obvious way to implement recollection is to use the same memory search processes that are used in free recall, as proposed by Gillund and Shiffrin (1984) and partially implemented by Ratcliff et al. (1995). Gillund and Shiffrin suggested two different ways in which recall processes could supplement recognition. First, a standard memory search could be conducted using context as the memory cue. If the test item is sampled and recovered, then it is a recallable item and is endorsed as old. Obviously, this is a separate decision process than the familiarity process and therefore the recognition decision would be made following the logic of the dual-process signal detection (DPSD) model of Yonelinas (1994): If the item is recollected, it is identified as old with the highest confidence rating. Otherwise, the decision is made based on familiarity.

Following this recommendation, I implemented two different memory search recollection mechanisms in which LTM is either cued with context (*context-only memory search*) or with context and the test item (*context+item memory search*). The idea that subjects engage in a standard free recall memory search every time they engage a recollection process seems rather unparsimonious and psychologically implausible. Therefore, I limited the memory search by assuming that subjects attempt up to L_{max} sample and recovery attempts with the provided cue. If the test item is sampled and successfully recovered, it is recollected. If this memory search is unsuccessful, then a recognition decision is made using familiarity.

Gillund and Shiffrin (1984) also suggested an even more limited memory search using a single sample and recovery attempt. To implement this mechanism, I used the algorithm described above, setting $L_{max} = 1$. Depending on whether memory is probed with context or the test item and context, I refer to these mechanisms as *context-only sampling* and *context+item sampling* recollection mechanisms.

Using the recovery rule for recollection. An alternative, and somewhat simpler, recollection mechanism is to calculate a value that represents the quantity and quality of contextual details that come to mind in response to the test item. In the current version of fSAM, this would most logically be a function of the absolute strength of association between the test item and the current context, *S(j, context)*. One logical function that retains the theoretical link between recall and recognition that is present in the Gillund and Shiffrin (1984) model is the recovery rule from free recall:

$$r(j) = 1 - e^{-S(j,context)}. \quad (5)$$

This non-linear transformation results in a value between 0 and 1 that can be used to make a discrete recollection decision (as in the DPSD model) or as a continuous measure of recollective strength (as in the UVSD model). Within the DPSD framework, the recovery value can be viewed as the probability that the item will be recollected. This probability can then be compared to a stochastically determined threshold, and if it is higher than the threshold, the item is recollected and endorsed as old with high confidence. This process forms the basis of the *discrete context-only* recollection mechanism.

Within the UVSD framework, *r(j)* can be used as a measure of recollective strength than can be combined with the familiarity value to calculate a value for the

total memory strength of the test probe. The *continuous context-only* recollection mechanism combines familiarity and recollection strengths using Equation 6:

$$S(j) = W_f f(j) + W_r r(j), \quad (6)$$

where W_f and W_r are weighting parameters for familiarity and recollection, respectively. This total memory strength value is then compared to a criterion, as in the UVSD model of recognition (Wixted, 2007a).

As with the memory search processes, I also created alternative versions of the recovery rule-based processes by including the item as part of the memory probe. The *discrete context+item* and *continuous context+item* recollection mechanisms use the context+item form of the recovery rule from fSAM (Kimball et al., 2007) instead of the context only rule shown above.

Testing the recognition mechanisms

To test the viability of these mechanisms, I constructed 36 model variants by factorially combining the four familiarity mechanisms with nine types of recollection mechanisms (no recollection plus the eight mechanisms discussed above), as shown in Table 7. In Simulation 1, I evaluated the ability of each of these models to generate a prototypical ROC and *z*-ROC for confidence ratings in an episodic recognition memory task. In Simulation 2, I tested the ability of each of the model variants to account for the word frequency mirror effect in simulations of a prototypical word frequency item recognition experiment (Glanzer & Adams, 1990, Experiment 2). These simulations are described in turn in the remainder of this chapter.

Table 7. Factorial combination of familiarity and recollection mechanisms integrated into the fSAM model in Simulations 1 and 2.

Decision model	Recollection mechanism	Familiarity			
		Global		Local	
		Product	Sum	Product	Sum
Familiarity	None				
DPSD	Context-only search				
DPSD	Context+Item search				
DPSD	Context-only sampling				
DPSD	Context+Item sampling				
DPSD	Context-only recovery				
DPSD	Context+Item recovery				
UVSD	Context-only recovery				
UVSD	Context+Item recovery				

Simulation 1: Confidence-based ROCs in Item Recognition

In a typical episodic item recognition experiment participants are presented with a list of unrelated words, one word at a time, followed by a recognition test containing both studied and unstudied words. During the recognition test, participants are asked to rate their confidence that each test word was studied, often using a scale of 1-6 (e.g., Yonelinas, 1994). Participant responses are then used to generate ROC curves. When participants study moderately long lists (50-200 items) of unrelated words and are tested within the same experimental session, they typically exhibit d' values between 1.0 and 2.0 and produce ROC curves that are curvilinear in probability space and linear with a

slope around 0.8 in *z*-space (for extensive reviews, see Wixted, 2007a; Yonelinas & Parks, 2007).

Data to be modeled. Each model was tested by simulating a prototypical recognition memory experiment in which subjects study a lists of 100 words and are given a test comprised of 50 randomly selected study items and 50 unrelated lures. An artificial data set representing typical results from an item recognition experiment was constructed for these simulations (see Table 8). These data produce a probability space ROC that is asymmetrical and a *z*-ROC with a slope of 0.8 and an intercept of 1.25, as shown in Figure 12.

Lexicon. A total of 500 words to be used as stimuli were randomly selected from the 5018 words from the University of Southern Florida word association norms (Nelson et al., 2004) that were used by Steyvers et al. (2004) to create their word association space (WAS). The Steyvers et al. version of WAS forms the basis of the semantic matrix in fSAM. These 500 words were randomly divided into a set of 100 words to be studied, a set of 50 words to be used as lures on the recognition test, and a set of 350 words that were placed in the model's lexicon but were not studied or tested.

Fitting method. Because the models are highly complex, I used a genetic algorithm (Ashlock, 2006; Mitchell, 1996) to estimate the best fitting parameters for each model. The algorithm started by creating an initial generation of 200 parameter sets by randomly drawing the value of each parameter in a set from a predetermined range of values. The model was then run with 20 simulated subjects using each of these randomly generated parameter sets. The mean number of targets and lures that were

Table 8. Data to be fit for Simulation 1. Data represent mean proportions of targets and lures endorsed at each level of confidence.

Confidence		Lure	Target
Sure old	1	0.06	0.47
	2	0.06	0.16
	3	0.10	0.12
	4	0.16	0.11
	5	0.30	0.08
Sure new	6	0.32	0.06

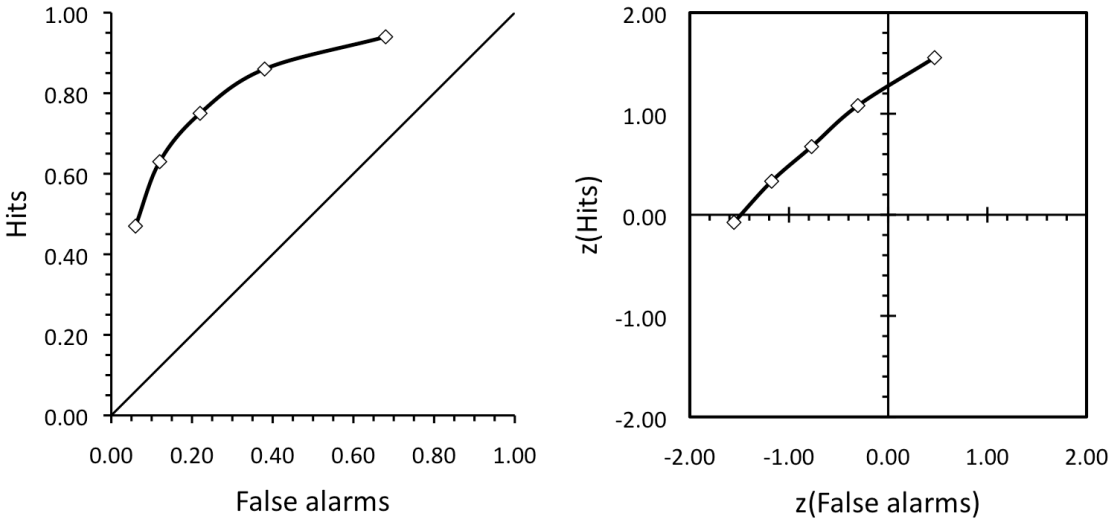

Figure 12. Confidence ROC and *z*-ROC plots for data to be fit in Simulation 1.

rated at each level of confidence was calculated across the simulated subjects and compared to the values shown in Table 8. Goodness of fit was calculated using unweighted root mean squared deviations (RMSD).

The first-generation parameter sets that yielded the lowest RMSD values were then used to create the next generation of parameter sets through the processes of mutation and recombination. Mutation creates a particular second-generation parameter set by randomly selecting one of the best fitting first generation parameter sets and then randomly copying or varying the value of each parameter within a specified range. During mutation, parameters were allowed to mutate by up to 1% of the parameter's range, in either direction, subject to the parameter's upper and lower range limits. Recombination creates a second-generation parameter set by randomly selecting two of the best fitting first-generation parameter sets as "parents" and, for each parameter, randomly selecting one parent's values for that parameter as the "child's" value.

For each generation, the 10 best-fitting parameter sets from the previous generation were kept, the next 20 best-fitting sets were mutated, and the remaining 170 sets were replaced with new parameter sets created by randomly recombining the 10 best-fitting sets and the 20 mutated sets. These processes iterated through 30 generations. The 100 best-fitting parameter sets generated by the genetic algorithm were then each used to run the model again using samples of 200 simulated subjects to generate statistically stable predictions and correct for regression to the mean. The parameter set from this large-sample run that had the lowest RMSD was used as the base parameter set for that model.

Results. Most of the models were able to provide a good quantitative fit to the raw confidence rating proportions. Table 9 shows the RMSDs for the best fits of each model to the data. The best fitting model was the version that combined the global product familiarity mechanism with the continuous (UVSD) context+item recovery recollection mechanism (RMSD = 0.010), but a number of other models provided fits that were nearly as good. Notably, the two versions that used local familiarity rules without any recollection mechanism were among the best fitting models. In no case did adding a recollection mechanism to either of the local familiarity mechanisms improve the fit, and it actually harmed the fit in most cases.

Table 9. Goodness of fit (RMSD) to confidence-based ROCs in Simulation 1.

Decision model	Recollection mechanism	Familiarity			
		Global		Local	
		Product	Sum	Product	Sum
Familiarity	None	0.019	0.091	0.011	0.012
DPSD	Context-only search	0.012	0.073	0.013	0.012
DPSD	Context+Item search	0.015	0.021	0.014	0.013
DPSD	Context-only sampling	0.012	0.076	0.020	0.013
DPSD	Context+Item sampling	0.010	0.041	0.013	0.013
DPSD	Context-only recovery	0.033	0.036	0.026	0.019
DPSD	Context+Item recovery	0.026	0.039	0.016	0.019
UVSD	Context-only recovery	0.016	0.014	0.014	0.013
UVSD	Context+Item recovery	0.010	0.017	0.012	0.014

Even though many of the models fit well quantitatively, most of the models did not do so well in capturing the shape of the confidence ROC and *z*-ROC. Tables 10 and 11 show the *z*-ROC slopes and intercepts, respectively, produced by each model. An examination of these data reveals a few interesting patterns. First, most of the models were not able to match the slope of the *z*-ROC, generating normalized slopes close to—and in some cases greater than—unity. This was particularly true for the models that used the global product familiarity mechanism based on the SAM model (Gillund & Shiffrin, 1984) or the global sum familiarity mechanism. This is consistent with previous studies in which SAM and other GMMs have been shown to predict ROCs with a slope greater than or equal to 1.0 (Clark & Gronlund, 1996; Ratcliff et al., 1992), suggesting that this is a problem for all strength-based global familiarity mechanisms.

By contrast, models based on a local familiarity mechanism did a much better job of fitting the normalized slope. In particular, the local product familiarity mechanism performed quite well, either alone or when combined with a recollection mechanism. The local product familiarity model (with no recollection) not only fit the normalized slope perfectly; it also fit the intercept quite well. This qualitative fit was improved slightly by adding either the DPSD context-only sampling recollection mechanism or the UVSD continuous context+item recovery rule recollection mechanism. Although local familiarity mechanisms are known to have problems predicting list length effects (Ratcliff et al., 1994), they appear to do quite well in predicting normalized ROC slopes.

Table 10. z-ROC slope for each model tested in Simulation 1. The data that were fit had a slope of 0.80. Good qualitative fits (0.70, 0.90) are in bold.

Decision model	Recollection mechanism	Familiarity			
		Global		Local	
		Product	Sum	Product	Sum
Familiarity	None	0.95	1.10	**0.80**	0.91
DPSD	Context-only search	1.00	1.10	**0.85**	0.92
DPSD	Context+Item search	0.93	**0.79**	0.92	0.93
DPSD	Context-only sampling	1.04	1.09	**0.79**	**0.86**
DPSD	Context+Item sampling	**0.89**	**0.72**	0.99	**0.84**
DPSD	Context-only recovery	0.62	0.65	1.01	1.05
DPSD	Context+Item recovery	0.94	0.59	0.93	1.07
UVSD	Context-only recovery	0.97	0.90	0.92	0.95
UVSD	Context+Item recovery	0.90	0.97	**0.80**	**0.85**

Table 11. z-ROC intercept for each model tested in Simulation 1. The data that were fit had a slope of 1.25. Good qualitative fits (1.15, 1.35) are in bold.

Decision model	Recollection mechanism	Familiarity			
		Global		Local	
		Product	Sum	Product	Sum
Familiarity	None	**1.29**	0.29	**1.33**	1.37
DPSD	Context-only search	1.43	0.52	**1.28**	1.43
DPSD	Context+Item search	1.41	**1.19**	**1.34**	1.42
DPSD	Context-only sampling	1.52	0.53	**1.22**	1.44
DPSD	Context+Item sampling	**1.34**	0.81	1.48	**1.29**
DPSD	Context-only recovery	**1.20**	0.89	1.65	1.61
DPSD	Context+Item recovery	1.52	0.90	1.45	1.63
UVSD	Context-only recovery	1.44	**1.29**	1.39	1.47
UVSD	Context+Item recovery	**1.35**	1.47	**1.28**	1.41

Simulation 2: Word Frequency Mirror Effects

Having shown in Simulation 1 that a strength-based memory model can generate ROCs that approximate ROCs created with data from human subjects, I next tested the ability of these models to simulate the word frequency mirror effect. For this test, I simulated a prototypical word frequency item recognition experiment (Glanzer & Adams, 1990, Experiment 2). In this experiment, subjects were presented with a study list of 148 words (50 high frequency targets, HF; 50 low frequency targets, LF; 24 HF fillers; and 24 LF fillers) one at a time for 1s each. They were then given an old-new recognition test with confidence ratings. The test was comprised of the 50 HF words, 50 LF words, and 100 unstudied lures (50 HF and 50 LF). The standard mirror effect was obtained, with LF words having a higher HR and lower FAR than the HF words.

Using the best-fitting parameter sets from Simulation 1, I used each model to simulate the procedure used by Glanzer and Adams (1990, Experiment 2) with 200 subjects. Importantly, unlike most other attempts to model the WFME, I made no *a priori* assumptions regarding the characteristics of LF and HF words, nor did I use different encoding parameters or retrieval strategies for the different classes of stimuli (cf., Glanzer et al., 1993; Reder et al., 2000; Shiffrin & Steyvers, 1997). The goal of this simulation was to capture the WFME as a natural consequence of differences between LF and HF words that are captured by the WAS metric (Monaco et al., 2007).

As Table 9 clearly shows, none of the models were able to produce a mirror effect with this set of assumptions. Although a few of the global familiarity models were able to capture the lower false alarm rate for LF words relative to HF words, none of them were able to capture the higher hit rate for LF words. In fact, all of the global

Table 12. Mean hit rates (HR) and false alarm rates (FAR) for high-frequency (HF) and low-frequency (LF) words in Simulation 2

Familiarity mechanism	Decision model	Recollection mechanism	RMSD	High frequency		Low frequency		Differences		Mirror effect?
				FAR	HR	FAR	HR	FAR	HR	
Empirical data from Glanzer and Adams (1990, Experiment 2)				0.28	0.61	0.23	0.70	-0.05	0.09	YES
Global product	Familiarity	None	0.10	0.39	0.71	0.34	0.60	-0.05	-0.11	NO
Global product	DPSD	Context-only search	0.08	0.37	0.72	0.27	0.64	-0.10	-0.08	NO
Global product	DPSD	Context+Item search	0.15	0.44	0.70	0.45	0.61	0.01	-0.10	NO
Global product	DPSD	Context-only sampling	0.09	0.38	0.74	0.31	0.65	-0.08	-0.09	NO
Global product	DPSD	Context+Item sampling	0.14	0.42	0.75	0.43	0.66	0.00	-0.09	NO
Global product	DPSD	Context-only recovery	0.11	0.37	0.70	0.38	0.58	0.01	-0.11	NO
Global product	DPSD	Context+Item recovery	0.09	0.32	0.76	0.32	0.65	0.00	-0.11	NO
Global product	UVSD	Context-only recovery	0.09	0.35	0.76	0.32	0.72	-0.03	-0.03	NO
Global product	UVSD	Context+Item recovery	0.09	0.34	0.75	0.32	0.71	-0.02	-0.04	NO
Global sum	Familiarity	None	0.24	0.58	0.62	0.52	0.46	-0.06	-0.16	NO
Global sum	DPSD	Context-only search	0.25	0.60	0.66	0.59	0.57	-0.01	-0.09	NO
Global sum	DPSD	Context+Item search	0.09	0.30	0.61	0.33	0.54	0.03	-0.07	NO
Global sum	DPSD	Context-only sampling	0.27	0.64	0.68	0.63	0.62	-0.02	-0.06	NO
Global sum	DPSD	Context+Item sampling	0.14	0.29	0.49	0.29	0.46	0.01	-0.04	NO
Global sum	DPSD	Context-only recovery	0.11	0.35	0.63	0.36	0.55	0.01	-0.08	NO
Global sum	DPSD	Context+Item recovery	0.08	0.35	0.65	0.33	0.60	-0.02	-0.05	NO
Global sum	UVSD	Context-only recovery	0.06	0.31	0.67	0.29	0.61	-0.03	-0.05	NO
Global sum	UVSD	Context+Item recovery	0.09	0.35	0.75	0.33	0.72	-0.02	-0.04	NO

Continued on next page

Table 12. continued

Familiarity mechanism	Decision model	Recollection mechanism	RMSD	High frequency FAR	High frequency HR	Low frequency FAR	Low frequency HR	Differences FAR	Differences HR	Mirror effect?
Empirical data from Glanzer and Adams (1990, Experiment 2)				0.28	0.61	0.23	0.70	-0.05	0.09	YES
Local product	Familiarity	None	0.07	0.32	0.71	0.32	0.72	-0.00	0.00	NO
Local product	DPSD	Context-only search	0.07	0.32	0.71	0.32	0.69	0.00	-0.01	NO
Local product	DPSD	Context+Item search	0.06	0.31	0.71	0.30	0.69	-0.01	-0.02	NO
Local product	DPSD	Context-only sampling	0.05	0.30	0.68	0.30	0.68	-0.00	0.00	NO
Local product	DPSD	Context+Item sampling	0.06	0.29	0.70	0.29	0.70	-0.00	-0.00	NO
Local product	DPSD	Context-only recovery	0.10	0.32	0.78	0.32	0.76	0.00	-0.02	NO
Local product	DPSD	Context+Item recovery	0.06	0.29	0.72	0.29	0.69	-0.00	-0.03	NO
Local product	UVSD	Context-only recovery	0.06	0.32	0.71	0.30	0.69	-0.02	-0.01	NO
Local product	UVSD	Context+Item recovery	0.07	0.33	0.71	0.32	0.70	-0.00	-0.01	NO
Local sum	Familiarity	None	0.07	0.31	0.72	0.31	0.72	-0.01	-0.01	NO
Local sum	DPSD	Context-only search	0.07	0.32	0.72	0.32	0.73	-0.00	0.01	NO
Local sum	DPSD	Context+Item search	0.06	0.31	0.70	0.30	0.69	-0.01	-0.00	NO
Local sum	DPSD	Context-only sampling	0.08	0.31	0.75	0.30	0.72	-0.01	-0.03	NO
Local sum	DPSD	Context+Item sampling	0.05	0.26	0.71	0.27	0.69	0.00	-0.01	NO
Local sum	DPSD	Context-only recovery	0.09	0.30	0.76	0.30	0.74	-0.00	-0.01	NO
Local sum	DPSD	Context+Item recovery	0.10	0.33	0.77	0.31	0.74	-0.02	-0.03	NO
Local sum	UVSD	Context-only recovery	0.06	0.28	0.72	0.28	0.72	0.00	-0.01	NO
Local sum	UVSD	Context+Item recovery	0.07	0.31	0.72	0.30	0.71	-0.01	-0.01	NO

familiarity models predicted *lower* hit rates for LF words than for HF words. The results from the local familiarity models were just as bad: None of the local familiarity models were able to capture any substantial effects of word frequency. These findings are somewhat surprising given Monaco et al.'s (2007) results showing that using WAS, LF words generate a higher familiarity signal than do HF words. Clearly, none of the recognition mechanisms I incorporated into the fSAM model are able to take advantage of this factor.

In order to try to understand why every variant of the fSAM model that I had tried failed to generate higher hit rates for LF than HF words, I examined the familiarity and recollection rules in more detail, and I found three fatal weaknesses in the SAM framework that militate against developing it further for use with recognition. First, I found that one of the most important mechanisms in the fSAM model—the semantic encoding mechanism that increases contextual associations during study as a function of association strengths in WAS—is negated by the use of a global familiarity rule. Because the global familiarity rules (Equations 1 and 2) sum the contextual associations over all items in memory, contextual strength can be factored out and will be the same for any cue set. Therefore, as Gillund and Shiffrin (1984) put it, context "acts as a scale factor in recognition and does not affect performance" (pg. 19). As a consequence, differences in contextual association strength between LF and HF words due to operation of the semantic encoding mechanism in fSAM have no effect on global familiarity. (This also implies that an fSAM recognition model using a global familiarity rule would not be able to account for semantically induced false recognition in the DRM paradigm, a key goal of the effort to extend fSAM to recognition.)

Second, although the local familiarity rules (Equations 3 and 4) do not have the problems with context that the global familiarity rules do, they do not take advantage of semantic associations during retrieval. In WAS, an item's self-similarity is always 1; therefore, the $S_s(j,j)$ term can be eliminated from the local product rule (Equation 3) and becomes a constant in the local sum familiarity rule (Equation 4). Thus, a model with either of these rules will have difficultly handling test-based effects involving semantic variables, such as the test-facilitated false memory observed in Experiment 1 and in Kimball et al. (2010).

Finally, it turns out that the recollection processes all have the same limitations as the fSAM recall model (Kimball et al., 2007) in that they can increase endorsements of items as old, but without some kind of monitoring process they cannot reduce endorsements. This is not a problem if recollection is only used to endorse items, as is assumed by the DPSD model (Yonelinas, 1994) and the recollection hypothesis for mirror effects (Joordens & Hockley, 2000). But if recollection is also used to reject items, as proposed by Brainerd et al. (2001), then these mechanisms are incomplete.

In summary, I developed and tested 36 different variants of an fSAM recognition model that factorially combined global and local familiarity mechanisms with a variety of plausible recollection mechanisms. Of these 36 models, 6 were able to account for confidence-based ROCs, including the *z*-ROC slope, but none were able to account for the WFME. Detailed examinations of the familiarity and recollection rules revealed that this failure was due to serious structure flaws in the SAM framework such that none of the mechanisms are able to take full advantage of semantic encoding and retrieval mechanisms in the fSAM model.

Chapter 8

Conclusions

The title of this thesis poses the question of whether recollection can explain mirror effects in item recognition. I used a two-prong approach to investigate this question. First, I examined the effects of normative word frequency and semantic relatedness on true and false recognition in 2 experiments with human participants. In Experiment 1, I observed the standard word frequency mirror effect for comparisons of LF words to HF words, but the effect did not obtain for comparisons of MF, HF, or VHF words. Additionally, there was no mirror effect for semantic relatedness. These findings run counter to predictions from Bayesian likelihood models that, absent a systematic bias, any variable that affects recognition performance will produce a mirror effect. The hypothesis that mirror effects are the result of differences in recollectability was partially supported by analyses of ROCs using the UVSD and DPSD measurement models. Experiment 2 showed further support for this hypothesis: The patterns for hit rates and "remember" judgments for targets were almost identical, as would be expected if the hit rate portion of the WFME is due to recollection. Importantly, the false alarm rates to non-critical lures showed effects of word frequency, but there were no such effects for remember judgments for lures. Together with previous studies, these experiments show the importance of examining mirror effects in more detail and considering boundary conditions for mirror effects when evaluating theories of memory (Greene, 2007).

The second part of this thesis was to try to develop a computational model of recognition memory that could parsimoniously account for mirror effects and their absence, as well as effects of semantic relatedness, in item recognition. As discussed in Chapters 3 and 6, strength-based familiarity global memory models—including previous implementations of the SAM model—underpredict mirror effects, while the Bayesian likelihood models that have largely replaced GMMs overpredict mirror effects. The SAC model developed by Reder and colleagues (Cary & Reder, 2003; Reder et al., 2000) is able to account for word-frequency mirror effects as well as some boundary conditions on these effects, but it cannot account for effects of semantic relatedness such as false memory in the DRM paradigm. Therefore, a "Goldilocks" model that can simultaneously handle mirror effects, boundary conditions, and semantic relatedness effects is highly desirable.

I tried to develop such a model by modifying the fSAM recall model (Kimball et al., 2007). I tested a total of 36 variants of the model that factorially combined 4 different familiarity mechanisms with a variety of recollection mechanisms. Some of these models used the DPSD decision process logic in which recognition was based on separate recollection and familiarity processes. Others used a mixture decision process in which familiarity and recollective strength were combined into a unidimensional UVSD decision model. Although some of these models were able to generate properly shaped ROCs and *z*-ROCS—something no other SAM models have been able to do—none of them were able to generate a WFME. Detailed investigations of the operation of the familiarity and recollection rules revealed that this was due to critical failures in the fSAM framework. These failures indicate the limits of the usefulness of the SAM

framework and point to key characteristics that a recognition model will need in order to fully account for mirror effects and their boundary conditions.

Model Failures and a Way Forward

The most problematic aspect of the SAM framework is the representation of context. When Gillund and Shiffrin (1984) developed the episodic SAM recognition model, they chose an architecture that would ensure that the familiarity of different items was not affected by the contextual strength of those items; that is, familiarity judgments can be considered to be context-free. This is seemingly consistent with the idea that familiarity is a global measure of the strength of the representation of a stimulus in memory that does not depend on the ability to remember or identify contextual details about previous encounters with that stimulus (e.g., Jacoby, 1991; Mandler, 1980; Wixted, 2007a; Yonelinas, 1994). However, on closure examination, the global familiarity rule used in SAM turns out to be a very different conceptualization of familiarity than is commonly thought.

Most theories of memory that include a familiarity process assume that familiarity strength is a function of a sensory and perceptual experience along with interitem integration (e.g. Juola, 1973). The SAM model represents these two types of evidence in different associative structures: context and the episodic matrix, respectively. However, because context is essentially factored out when calculating familiarity, differences in prior sensory and perceptual experiences between items cannot contribute to the familiarity signal in SAM. This explains why the Gillund and Shiffrin (1984) SAM recognition model has never been able to successfully account for context-based phenomena (Clark & Gronlund, 1996). In order to account for these

types of effects, the SAM model would need 1) a richer contextual representation and 2) a new way of calculating familiarity that allows for different contextual strengths in response to different test probes.

Mensink and Raaijmakers (1988) provided SAM with a rich contextual representation by incorporating Estes' (1955) stimulus fluctuation and sampling theory into the SAM framework. In the Mensink and Raaijmakers SAM model, context is represented as a vector of elements in which each contextual element is in either an active or inactive state at any given time. The identities of the active contextual elements change over time, with some active elements becoming inactive and some inactive elements becoming active at each time step. Each time an item is studied, a proportion of the active contextual elements are conditioned, or associated, to the item on a binary basis. The conditioning rule is expressed as a differential equation so that the number of newly conditioned elements is a decreasing function of the number of currently conditioned elements, generating an exponential learning curve. The contextual retrieval strength of an item in memory is determined by the number of contextual elements active at the time of test that were associated to the item during study. These changes allowed Mensink and Raaijmakers to account for a number of interference phenomena in cued recall and forced choice recognition (using a recall process).

Unfortunately, though, the Mensink and Raaijmakers (1988) SAM model has never been applied to item recognition, and even if it were to be applied to item recognition it would still have the same problem as the base SAM model because differences in context between items would still not affect familiarity. It might be

possible to use the Mensink and Raaijmakers model with a local familiarity rule (as in Simulation 1) rather than a global rule, but the model would still be limited to episodic effects and would not be able to account for semantic effects in recognition. Kimball et al. (2007) proposed that the contextual fluctuation mechanism could be implemented in fSAM recall model, thereby creating a model that could account for both contextual and semantic influences in recall. Several years ago I attempted to do this, but I was unable to find a way to reconcile the vector representations used in the contextual fluctuation mechanism with the associative matrix representation of semantic memory.

Based on these experiences, I have reluctantly concluded that even with the modifications that have been made to the SAM framework over the years, it is not suitable as a basis for a general model of recall and recognition. I now briefly outline a new model that keeps the most desirable features from fSAM—namely the use of a preexperimental semantic memory, an episodic-semantic encoding mechanism, and a conjunctive memory search algorithm—and places these in an architecture that is more suitable for modeling familiarity and recollection.

An alternative to the association-based representations used in fSAM is to use a featural representation for items. Most other memory models, including TODAM2 (Murdock, 1993), MINERVA 2 (Hintzman, 1988), and REM (Shiffrin & Steyvers, 1997) use featural representations. A featural representation has a number of advantages over an association-based representation. First, it gives the model a way to represent similarities and differences between individual stimuli; stimuli that share more features are, by definition, more similar. This capability is critical in both familiarity assessments and recollection and is likely to be important in modeling stimulus-based

mirror effects such as the word-frequency effect (e.g., Shiffrin & Steyvers, 1997). Second, a featural representation gives the model a rich way of representing environmental and mental context, another necessity for recollection. Finally, featural representations can help bridge the gap between high-level models of cognitive processes and lower-level models such as neural networks and neurobiological models (Howard & Kahana, 2002).

Despite the advantages of featural representations, the association-based structure of the SAM framework should not be completely abandoned. Associations have been—and will continue to be—an important part of memory theory, and it seems to me that a general model of recall and recognition ought to include associations as a core component of the model. Therefore, the challenge is in how to combine the two types of representations (associations and features) in a way that is both parsimonious and psychologically plausible.

One possible solution is to think of memory as consisting of a set of linked neural networks in which information about item features is generated, encoded, and stored. This information can be computationally represented as sets of abstract feature values that are mathematically equivalent to locations in a high-dimensional space (e.g, Howard & Kahana, 2002) so that an item is represented by a location in memory space. Although associations are not stored directly in long-term memory in this architecture, they nevertheless can play key roles in memory processes because associations are a natural byproduct of feature overlap. That is, the strength of association between any two items in memory can be calculated as a function of the relative locations of those items in memory space. Items that are close together in memory space will tend to be

highly associated with each other, while items that are far apart in memory space will not.

Deriving associations within a spatial representation is, of course, not entirely novel. Indeed, this is the central idea behind data analysis techniques such as factor analysis and multi-dimensional scaling (Kruskal & Wish, 1978), lexical-semantic tools such as word association space (Steyvers et al., 2005) and latent semantic analysis (Landauer & Dumais, 1997), and some models of similarity judgment (Krumhansl, 1978; Tversky, 1977). Additionally, the temporal context model of memory and its derivatives (Howard & Kahana, 2002; Polyn, Norman, & Kahana, 2009) are based on this type of representation.

However, I believe that combining a spatial-association representation with the encoding and retrieval processes that have proven to be highly successful within the SAM framework is an exceptionally promising avenue to explore. In particular, using a featural representation of both items and context would provide a basis for implementing recollection and source monitoring mechanisms within the model. As discussed in Chapter 3 and suggested by the results of Experiments 1 and 2, recollection appears to play a key role in mirror effects and their boundary conditions. These mechanisms are also likely to be important in a number of other paradigms, including episodic recognition, false recall, and false recognition. Additionally, because such a model would include representations of both items and associations, it could provide a good platform from which to explore dissociations between item recognition and associative recognition—a task that has proven to be challenging for even the most complex computational models of memory (e.g., Malmberg, 2008).

Of course, there are a number of challenges to overcome in the development of such a model. One of the most critical issues is the fact that similarity and association are different constructs, even though many models—including fSAM—do not really distinguish between them (e.g., Kimball et al., 2007; Shiffrin & Steyvers, 1997; Murdock, 1993). Together with my advisor, Dan Kimball, I am currently working on finding a way to derive separate measures of similarity and association from the same underlying featural representations. One possibility is to use angular distance as the measure of association as many current models do, while using a metric such as Krumhansl's (1978) distance-density model as a measure of similarity. Despite the difficulties of this task, I remain hopeful.

References

Anaki, D., Faran, Y., Ben-Shalom, D., & Henik, A. (2005). The false memory and the mirror effects: The role of familiarity and backward association in creating false recollections. *Journal of Memory and Language, 52(1)*, 87-102.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*(2), 97-123.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Oxford, England: V. H. Winston & Sons.

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory & Cognition, 2*(3), 406-412.

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language, 39*(3), 371-391.

Arndt, J., & Reder, L. M. (2002). Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 830-842.

Ashlock, D. (2006). *Evolutionary computation for modeling and optimization* New York: Springer.

Atkinson, R. C. (1963). A variable sensitivity theory of signal detection. *Psychological Review, 70(1)*, 91-106.

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language, 46*, 199-226.

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review, 110*(4), 762-784.

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 307-327.

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and

improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977-990.

Buchler, N. G., Light, L. L., & Reder, L. M. (2008). Memory for items and associations: Distinct representations and processes in associative recognition. *Journal of Memory and Language, 59*(2), 183-199.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*, 231-248.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition, 20*(3), 231-243.

Clark, S. E. (1995). The generation effect and the modeling of associations in memory. *Memory & Cognition, 23*(4), 442-455.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*(1), 37-60.

Clark, S. E., & Shiffrin, R. M. (1992). Cuing effects and associative information in recognition memory. *Memory & Cognition, 20*(5), 580-598.

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember–know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review, 15*, 906-926.

Collins, A.M., & Loftus, E.F. (1975). A spreading-activation theory of semantic memory. *Psychological Review, 82,* 407-428.

Curran, T., & Hintzman, D. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(3)*, 531-547.

DeCarlo, L. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 18-33.

Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. *Psychological Reports, 5*, 305-312.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452-478.

de Zubicaraya, G. I., McMahona, K. L., Eastburna, M. M., Finnigana, S., & Humphreys, M. S. (2005). fMRI evidence of word frequency and strength effects in recognition memory. *Cognitive Brain Research, 24*, 587–598.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review, 13*(1), 1-21.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 414-435.

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

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*, 145-154.

Estes, W. K., & Maddox, W. T. (2002). On the processes underlying stimulus-familiarity effects in recognition of words and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1003-1018.

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

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67.

Glanzer, M., & Adams, J. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8-20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 5-16.

Glanzer, M., Adams, J., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 81-93.

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

Glanzer, M., Hilford, A., & Maloney, L.T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*, 431-455.

Green, D. (1960). Psychoacoustics and detection theory. *Journal of the Acoustical Society of America*, *32*1189-1203.

Greene, R. L. (2007). Foxes, Hedgehogs, and Mirror Effects: The Role of General Principles in Memory Research. In J. S. Nairne (Ed.), *The foundations of*

remembering: Essays in honor of Henry L. Roediger, III. (pp. 53-66). New York, NY: Psychology Press.

Greene, R. L., & Tussing, A. A. (2001). Similarity and associative recognition. *Journal of Memory and Language, 45*(4), 573-584.

Gregg, V. (1976). Word frequency, recognition and recall. In J. L. G. Brown (Ed.), *Recall and Recognition* (pp. 183-216). Oxford, England.

Heathcote, A., Ditton, E., & Mitchell, K. (2006). Word frequency and word likeness mirror effects in episodic recognition memory. *Memory & Cognition, 34(4)*, 826-838.

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language, 55*(4), 495-514.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528-551.

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 201-205.

Hintzman, D., Caulton, D., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 275-289.

Hirshman, E., Fisher, J., Henthorn, T., Arndt, J., & Passannante, A. (2002). Midazolam amnesia and dual-process models of the word-frequency mirror effect. *Journal of Memory and Language*, *47*(4), 499-516.

Hockley, W. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition*, *22*(6), 713-722.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269-299.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513-541.

Jacoby, L. L. (1998). Invariance in automatic influences of memory: Toward a user's guide for the process-dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(1),* 3-26.

Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1534-1555.

Juola, J. (1973). Repetition and laterality effects on recognition memory for words and pictures. *Memory & Cognition*, *1*(2), 183-192.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition, 24*(1), 103-109.

Kelley, C. M., & Jacoby, L. L. (2000). Recollection and familiarity: Process-dissociation. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 215-228). NY: Oxford University Press.

Kim, K., & Glanzer, M. (1994). Attention/likelihood: Reply to Hintzman (1994). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(1), 206-208.

Kimball, D.R., & Smith, T.A. (2007, November). *Generalizing the fSAM model: Simulation of core false recall effects*. Paper presented at the Annual Meeting of the Psychonomic Society, Long Beach, CA.

Kimball, D.R., Muntean, W. J., & Smith, T.A. (2010). Dynamics of thematic activation in recognition testing. *Psychonomic Bulletin & Review*, *17*, 355-361.

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review, 114*(4), 954-993.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III.* (pp. 171-189). New York, NY: Psychology Press.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*(5), 445-463.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills: Sage.

Kučera, H., & Francis, W.N., (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Landauer, T. K., & Dumais, S. T. (1997). Solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lehman, M., & Malmberg, K.J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 970-988.

Luce, D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70(1)*, 61-79.

Macmillan, N. A. (2002). Signal detection theory. In H. Pashler and J. Wixted (Eds.) *Stevens' Handbook of Experimental Psychology, Third Edition, Vol. 4: Methodology in Experimental Psychology* (pp. 43-90). New York: Wiley.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*.

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

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*(4), 335-384.

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 319-331.

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 616-630.

Malmberg, K. J., & Shiffrin, R. M. (2005). The 'one-shot' hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 322-336.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30*(4), 607-613.

Mandler, G. (1969). Input variables and output strategies in free recall of categorized lists. *American Journal of Psychology, 82(4)*(4), 531-539.

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

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724-760.

Mensink, G. J. M., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review, 95*(4), 434-455.

Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.

Monaco, J., Abbott, L., & Kahana, M. (2007). Lexico-semantic structure and the word-frequency effect in recognition memory. *Learning & Memory*, *14*(3), 204-213.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100*, 183-203.

Neely, J. H., & Tse, C.-S. (2007). Semantic Relatedness Effects on True and False Memories in Episodic Recognition: A Methodological and Empirical Review. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III.* (pp. 313-351). New York, NY: Psychology Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments & Computers, 36*(3), 402-407.

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review, 114*(1), 188-201.

Polyn, S.M., Norman, K.A., & Kahana, M.J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116,* 129-156.

Raaijmakers, J. G. W. (2005). Modeling Implicit and Explicit Memory. In C. Izawa & N. Ohta (Eds.), *Human learning and memory: Advances in theory and application: The 4th Tsukuba International Conference on Memory.* (pp. 85-105). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, pp. 207-262). New York.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Order effects in recall. In J. B. Long & B. A. D. (Eds.), *Attention and Performance IX* (pp. 403-415). Hillsdale, NJ: Erlbaum.

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

Rao, K., & Proctor, R. (1984). Study-phase processing and the word frequency effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 386-394.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 163-178.

Ratcliff, R. & McKoon, G. (2000). Memory models. In E. Tulving & F. I. M. Craik (Eds.) *The Oxford handbook of memory* (pp. 571-581). NY: Oxford University Press.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 763-785.

Ratcliff, R., Sheu, C.-f., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518-535.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General, 124*(4), 352-374.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(2), 294-320.

Roediger, H. L., III, Balota, D. A., & Watson, J. M. (2001). Spreading activation and arousal of false memories. In H. L. Roediger, III, J. S. Nairne, I. Neath & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95-115). Washington, DC: American Psychological Association.

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

Roediger, H. L., III, Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*(3), 385-407.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review, 111*(3), 588-616.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review, 12*(5), 865-873.

Seamon, J. G., Lee, I. A., Toner, S. K., Wheeler, R. H., Goodkind, M. S., & Birch, A. D. (2002). Thinking of critical words during study is unnecessary for false memory in the Deese, Roediger, and McDermott procedure. *Psychological Science, 13*(6), 526-531.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 267-287.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179-195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM--retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145-166.

Shepard, R. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning & Verbal Behavior*, *6*(1), 156-163.

Sirotin, Y. B., Kimball, D. R., & Kahana, M. J. (2005). Going beyond a single list: Modeling the effects of prior experience on episodic free recall. *Psychonomic Bulletin & Review, 12*(5), 787-805.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition, 33*(1), 151-170.

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1499-1517.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6*(4), 342-353.

Smith, T.A., & Kimball, D.R. (2008, November). *Modeling part-set cuing of false memories.* Poster presented at the Annual Meeting of the Psychonomic Society, Chicago, IL.

Smith, T. A., & Kimball, D. R. (in press). Pursuing a general model of recall and recognition. In A.S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork*. New York: Psychology Press.

Stadler, M. A., Roediger, H. L., III, & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*, 494-500.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. Washington, DC: American Psychological Association.

Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1397-1410.

Stretch, V., & Wixted, J. T. (1998b). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1379-1396.

Tussing, A., & Greene, R. (2001). Effects of familiarity level and repetition on recognition accuracy. *The American Journal of Psychology*, *114*(1), 31-41.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26*(1), 1-12.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

Verde, M. F. (2009). The list-strength effect in recall: Relative-strength competition and retrieval inhibition may both contribute to forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 205-220.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A-sub(z), and A'. *Perception & Psychophysics, 68*(4), 643-654.

Wais, P. E., Mickes, L., & Wixted, J. T. (2008). Remember/know judgments probe degrees of recollection. *Journal of Cognitive Neuroscience, 20*(3), 400-405.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY, US: Oxford University Press.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 681-690.

Wixted, J. T. (2007a). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152-176.

Wixted, J. T. (2007b). Spotlighting the probative findings: Reply to Parks and Yonelinas (2007). *Psychological Review, 114*(1), 203-209.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*(4), 616-641.

Xu, J., & Malmberg, K. J. (2007). Modeling the effects of verbal and nonverbal pair strength on associative recognition. *Memory & Cognition, 35*(3), 526-544.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1341-1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25*(6), 747-763.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441-517.

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember-know, process dissociation, and receiver operating characteristic data. *Neuropsychology, 12*(3), 323-339.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800-832.

Appendix – Stimuli for Experiments 1 and 2

Associative Lists – Critical words are in bold

LOW FREQUENCY	Freq	BAS	FAS
SEA	59.84		
CORAL	2.37	0.11	0.00
OCTOPUS	1.94	0.10	0.00
DIVER	2.43	0.04	0.00
SERPENT	1.94	0.03	0.00
NAVIGATOR	2.69	0.03	0.00
NEPTUNE	2.67	0.02	0.00
BUG	20.94		
BEETLE	2.06	0.61	0.05
ROACH	2.65	0.33	0.15
LICE	2.14	0.24	0.00
MOTH	2.27	0.09	0.00
ANNOY	2.51	0.06	0.01
NAG	2.18	0.03	0.00
HORSE	92.88		
UNICORN	2.47	0.39	0.00
OATS	2.06	0.20	0.00
ZEBRA	2.51	0.18	0.00
BUGGY	2.49	0.15	0.00
TACK	2.12	0.04	0.00
TAME	2.73	0.02	0.00
WOOD	27.00		
LUMBER	2.47	0.59	0.00
TIMBER	2.49	0.41	0.00
STUMP	2.45	0.02	0.00
PLAQUE	2.08	0.02	0.00
FLUTE	2.12	0.01	0.00
CUPBOARD	2.49	0.01	0.00
MILK	42.53		
QUART	2.08	0.46	0.00
GALLON	2.27	0.29	0.00
SAUCER	2.75	0.09	0.00
YOGURT	2.27	0.05	0.00
JUG	2.63	0.04	0.00
ULCER	2.57	0.01	0.00
RAIN	48.90		
CLOUDY	2.16	0.28	0.00
PUDDLE	1.94	0.17	0.02
DEW	2.14	0.07	0.00
GLOOMY	2.41	0.05	0.00
TORNADO	2.55	0.01	0.00
RUST	2.49	0.01	0.00

MODERATE FREQUENCY	Freq	BAS	FAS
FIRE	215.49		
SMOKE	65.43	0.29	0.02
BURN	55.22	0.26	0.06
MATCH	49.43	0.24	0.02
CAMP	51.22	0.13	0.00
EMERGENCY	54.43	0.02	0.00
BOMB	53.65	0.01	0.00
POLICE	236.16		
SHERIFF	61.08	0.33	0.00
ARREST	59.55	0.24	0.02
DETECTIVE	61.12	0.15	0.00
GUARD	58.20	0.07	0.00
SERGEANT	62.94	0.04	0.00
JAIL	70.63	0.01	0.00
GOD	903.16		
CHURCH	69.67	0.31	0.04
HEAVEN	56.61	0.12	0.09
CROSS	55.04	0.09	0.00
HOLY	68.14	0.09	0.02
SPIRIT	49.35	0.07	0.00
SOUL	76.96	0.05	0.00
SMELL	83.14		
DIRTY	66.45	0.17	0.00
TASTE	51.31	0.14	0.18
NOSE	69.75	0.10	0.12
AWFUL	63.41	0.05	0.01
FRESH	54.51	0.03	0.00
ROSE	53.02	0.03	0.00
PICTURE	138.45		
CAMERA	57.00	0.44	0.05
ART	70.80	0.10	0.00
SCENE	74.65	0.08	0.00
FILM	65.25	0.06	0.00
WALL	70.69	0.05	0.04
COPY	52.27	0.01	0.00
STAR	81.35		
MOON	49.96	0.11	0.12
SPACE	66.06	0.08	0.02
HERO	49.84	0.04	0.00
TRACK	55.75	0.01	0.00
NORTH	63.88	0.01	0.00
SUN	69.67	0.01	0.09

Associative Lists, continued

HIGH FREQUENCY	Freq	BAS	FAS
ALIVE	154.47		
DEAD	448.98	0.39	0.55
BEING	485.90	0.04	0.00
FIVE	285.45	0.03	0.00
LIVE	344.59	0.02	0.02
DEATH	216.69	0.01	0.01
DIE	261.14	0.01	0.00
MOTHER	479.92		
FAMILY	354.25	0.05	0.01
WIFE	348.92	0.03	0.00
BROTHER	283.94	0.02	0.00
WOMAN	434.63	0.02	0.02
KID	339.20	0.02	0.00
MOM	430.39	0.01	0.06
PRETTY	392.22		
BEAUTIFUL	279.73	0.38	0.10
LOOKS	311.49	0.13	0.00
FACE	289.16	0.05	0.00
GIRLS	208.35	0.03	0.00
LADY	217.08	0.01	0.00
EYES	221.55	0.01	0.00
SAD	63.37		
HAPPY	333.20	0.62	0.63
ALONE	308.53	0.08	0.00
FUNNY	218.18	0.07	0.00
WORRY	287.02	0.02	0.00
HURT	246.35	0.01	0.00
HATE	214.59	0.01	0.00
STUDY	49.04		
CASE	282.41	0.09	0.00
COURSE	487.22	0.06	0.00
SCHOOL	333.12	0.06	0.09
THINKING	281.43	0.03	0.00
READ	241.22	0.02	0.07
HOURS	214.88	0.01	0.00
BALL	104.96		
PLAY	354.53	0.13	0.00
HIT	275.00	0.11	0.02
JACK	251.59	0.02	0.00
MISS	467.65	0.02	0.00
HAND	279.65	0.01	0.00
POINT	236.53	0.01	0.00

VERY HIGH FREQUENCY	Freq	BAS	FAS
NEED	1294.90		
WANT	2759.18	0.28	0.60
MUST	699.24	0.06	0.00
HELP	921.12	0.03	0.02
MORE	1298.59	0.01	0.00
CALL	861.39	0.01	0.00
GET	4583.76	0.01	0.00
WHAT	9842.45		
WHO	2222.94	0.22	0.18
WHERE	1830.22	0.06	0.11
HOW	3056.22	0.04	0.05
WHY	2248.76	0.03	0.13
THAN	738.80	0.02	0.00
WHEN	2034.10	0.02	0.08
OUT	3865.31		
IN	9773.41	0.94	0.80
WAY	1424.73	0.38	0.00
WITH	5048.33	0.06	0.00
TIME	1958.63	0.05	0.00
GOING	2123.29	0.03	0.00
LOOK	1947.27	0.02	0.00
NO	5971.55		
YES	1996.76	0.83	0.76
NOT	5424.96	0.25	0.00
MAYBE	926.45	0.07	0.01
NEVER	1362.55	0.05	0.03
PLEASE	1100.96	0.04	0.00
KNOW	5721.18	0.01	0.00
NOTHING	853.61		
EVERYTHING	654.88	0.34	0.08
SOMETHING	1500.16	0.31	0.16
ANYTHING	907.25	0.21	0.03
ALL	5161.86	0.06	0.03
THING	1088.67	0.01	0.00
JUST	4749.14	0.01	0.00
GO	3793.04		
STOP	707.27	0.61	0.54
COME	3140.98	0.57	0.13
LET	2419.24	0.23	0.00
WAIT	830.25	0.04	0.00
DO	6135.59	0.03	0.00
AWAY	730.90	0.02	0.02

Non-associative Lists

LOW FREQUENCY	Freq	MODERATE FREQUENCY	Freq
FUNCTIONAL	2.73	LOW	59.14
INDOORS	2.47	EATING	60.82
LOATHE	2.08	WRITING	55.92
CORNWALL	2.02	INNOCENT	54.51
NAW	2.08	BOYFRIEND	72.24
DUPLICATE	2.29	DECISION	55.06
OUTBREAK	1.94	CUP	51.65
PEEPING	2.51	OBVIOUSLY	60.43
FLOP	2.31	OURSELVES	52.47
SERMON	2.73	SERIOUSLY	64.12
BALE	2.04	ACTION	61.08
DILEMMA	2.61	WARM	52.14
LEAKS	2.20	TELLS	52.25
VOILA	2.53	PROMISED	62.14
FIREMEN	2.14	NOTE	53.55
POSED	2.20	WITNESS	51.39
WORKPLACE	2.43	NORMAL	70.37
RADIATOR	2.02	BREAKFAST	66.29
STIRRED	2.41	SAKE	64.16
SPICES	2.24	CONGRATULATIONS	70.90
SUSPENSE	2.24	SHARE	69.51
SMUGGLE	2.00	HALL	51.94
PLATINUM	2.41	INTEREST	50.94
INVESTIGATORS	2.69	COMPUTER	59.04
SCROLL	2.27	WORST	56.35
PERJURY	2.25	WILD	57.31
PRIMARILY	2.06	BAND	53.41
CRUMBS	2.55	LEG	56.51
STEED	2.55	TWICE	62.57
TRUTHFULLY	2.35	EVIL	73.16
COMMANDANT	2.57	HUNGRY	77.08
PUFFS	2.51	LOVES	72.45
CONSUMER	2.08	WINE	60.35
FRACTION	2.04	GOTTEN	54.27
SWAB	2.37	YEP	50.98
JUNKYARD	2.69	LISTENING	62.84

Non-associative Lists, continued

HIGH FREQUENCY	Freq	VERY HIGH FREQUENCY	Freq
NOBODY	266.65	TALK	855.00
ASK	483.14	WORK	798.02
UNTIL	302.47	AT	3217.10
SAYING	291.65	THESE	904.00
HAVING	289.25	ONLY	1083.71
REST	212.96	DOES	666.71
WATCH	330.02	FEEL	627.24
DIFFERENT	209.53	DAY	801.82
DEAR	223.43	HE	7637.20
LEAST	207.76	IF	3541.37
WITHOUT	354.65	EVER	709.22
TRYING	448.02	ME	9241.94
JOB	413.00	ALWAYS	655.25
FOUR	255.78	HERE	4525.25
FOUND	396.00	SIR	964.47
QUITE	202.59	SURE	1099.82
READY	387.80	GUY	762.61
TROUBLE	223.55	LOVE	1114.98
FORGET	277.06	MY	6762.73
MINUTE	377.49	SEE	2556.73
CAR	483.06	MONEY	640.76
CHANGE	240.35	NOW	3202.61
ONCE	344.88	OKAY	2006.27
BIT	235.04	AROUND	736.73
GONE	296.76	MUCH	973.25
CUT	229.76	NIGHT	866.04
DEAL	261.37	FOR	6895.10
GUESS	453.98	NAME	641.86
MYSELF	342.55	TELL	1724.49
HEART	244.18	SHOULD	1061.94
USED	344.14	FIRST	840.57
PROBABLY	280.84	OVER	1323.29
OFFICE	203.90	GIVE	1167.82
KNEW	368.96	BACK	2009.16
YEAR	277.92	THINGS	692.88
KNOWS	244.94	OR	1705.29

Buffer words for primacy and recency

	Freq
BANKS	15.90
BOWMAN	3.57
BROTHERS	47.06
BURNT	9.57
COLLECTING	6.84
COMPETITIVE	4.20
COOPERATE	10.35
COSY	2.02
DREAMT	6.65
FAKE	36.33
GAIN	13.73
HARBOR	11.02
KINDS	21.98
MUSEUM	18.47
PINNED	4.53
PROJECT	37.39
PUSH	70.55
REPORTS	25.41
SAFELY	11.10
SECTOR	7.06
SUITED	2.98
UNCLES	2.16
WAKE	105.22
WEDDING	101.43