EXAMINING THE TESTING EFFECT USING THE DUAL-PROCESS SIGNAL DETECTION MODEL

By

Nicole J. Bies-Hernandez

Bachelor of Science in Psychology
Fayetteville State University
2006

Master of Arts in Psychology
Fayetteville State University
2008

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy in Psychology

Department of Psychology
College of Liberal Arts
The Graduate College

University of Nevada, Las Vegas
May 2013

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend the dissertation prepared under our supervision by

Nicole J. Bies-Hernandez

entitled

Examining the Testing Effect Using the Dual-Process Signal Detection Model

be accepted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Psychology
Department of Psychology

David E. Copeland, Ph.D., Committee Chair

Mark H. Ashcraft, Ph.D., Committee Member

Joel S. Snyder, Ph.D., Committee Member

Colleen M. Parks, Ph.D., Committee Member

CarolAnne M. Kardash, Ph.D., Graduate College Representative

Tom Piechota, Ph.D., Interim Vice President for Research &
Dean of the Graduate College

May 2013

ABSTRACT

Taking a test can lead to enhanced long-term retention compared to not practicing the information or simply restudying, a finding known as the testing effect (Roediger, Agarwal, Kang, & Marsh, 2010). The current study examined whether the dual-process signal detection (DPSD) model (Yonelinas, 1994) offers an approach for investigating the testing effect across two experiments. Experiment 1 investigated if the DPSD model could be used to examine the testing effect, and it also examined a factor (i.e., the number of practice sessions) that influences the magnitude of the testing effect. Experiment 2 investigated whether making the final test dependent on recollection would influence the magnitude of the testing effect and the parameter estimates of recollection and familiarity. The results of these experiments demonstrated that when practice testing enhanced later memory, it also influenced the processes underlying the recognition memory judgments in a manner consistent with the DPSD model. Practice testing (in comparison to restudying) increased familiarity in both experiments and increased both familiarity and recollection when three practice tests were used. However, when comparing old versus similar lure items on the recollection-dependent final test format, no significant differences between practice testing and restudying were found. Overall, this study demonstrated that the DPSD model can be used to examine the testing effect. The DPSD model may provide a useful approach for future research investigating the testing effect in terms of the conditions under which the effect occurs, factors that influence the effect, and theoretical explanations for the effect.

ACKNOWLEDGEMENTS

I want to express my sincere gratitude and appreciation to everyone who contributed to the study undertaken for this dissertation as well as to both my personal growth and personal development more generally. First, I would like to acknowledge the significant support and scholarly advice from my dissertation committee: Dr. David Copeland, Dr. Mark Ashcraft, Dr. Colleen Parks, Dr. Joel Snyder, and Dr. CarolAnne Kardash. Each member's unique perspective and support of this project has been genuinely appreciated. I am also very greatful for the many experiences I have had learning and interacting with them, whether in classes, in the department, in the laboratory, or at conferences. I would also like to sincerely thank my advisor, Dr. David Copeland, as he has been a tremendous advisor and his continual support, advice, and encouragement have helped me to not only complete this dissertation project but more importantly, grow personally and professionally into an academic. Finally, I want to acknowledge the colleagues who have supported me throughout the process of completing this project and my graduate study at UNLV. Kris Gunawan and Katie Larson have been sources of support, constructive criticism, and friendship as part of the Reasoning and Memory Lab. I am also especially grateful to Erica Noles for her support and friendship.

DEDICATION

I wish to dedicate this dissertation to my beloved husband, Esidro Hernandez. Without his love, understanding, and encouragement I would not have persevered in accomplishing this dissertation nor pursuing my educational goal of earning a doctorate degree. I would also like to dedicate this dissertation to my parents, David and Susan Bies, and my brother, David M. Bies, who consistently loved and supported me throughout my long journey through post-baccalureate study.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Research has shown that testing leads to enhanced performance on a future test of long-term retention compared to restudying or not practicing the information (i.e., the testing effect; Roediger & Karpicke, 2006b). Very few studies have examined the testing effect using a dual-process or signal detection based approach. This dissertation examined the testing effect using the dual-process signal detection model (Yonelinas, 1994). First, the testing effect is explained, including factors that can affect it and theories that have been proposed to explain it. Then, the dual-process signal detection model is discussed in terms of the assumptions of the model and a prominent procedure used to assess the model (i.e., using receiver operating characteristics; Yonelinas & Parks, 2007). Finally, a set of experiments that investigated whether the dual-process signal detection model provides a useful approach for examining the testing effect are described, and the implications of these experiments are discussed.

CHAPTER 2

REVIEW OF RELATED LITERATURE

Testing Effect

There is a standard assumption that learning occurs during study (i.e., the encoding of information), while testing is simply a neutral way to assess learning without influencing it. Based on this standard assumption, in educational settings as well as other settings, tests are typically used to assess learning. However, contrary to the standard assumption just described, research has not supported the notion that testing is a neutral event that measures learning without affecting it (Roediger, Agarwal, Kang, & Marsh, 2010; Roediger & Karpicke, 2006b). Instead, it has been demonstrated that learning can occur during testing; this finding is referred to as *test-enhanced learning* (McDaniel, Roediger, & McDermott, 2007). More specifically, test-enhanced learning refers to the fact that retrieval (e.g., through testing) can have positive influences on learning and memory (McDaniel et al., 2007).

Research has shown that there are specific effects related to test-enhanced learning: direct and indirect effects. The direct effect, referred to as the *testing effect*, is based on the finding that when people are tested after initial encoding, testing leads to enhanced memory in the future compared to simply restudying the material (e.g., re-reading the material) or when the material is not practiced (Roediger et al., 2010; Roediger & Karpicke, 2006b). That is, the act of taking a test can directly increase long-term retention of the tested (or related) material (Roediger & Butler, 2011; Roediger & Karpicke, 2006b). The testing effect is a very robust effect, and it has been shown to occur for a variety of stimuli and experimental conditions. Generally, the evidence for the

testing effect demonstrates that testing is a powerful way to enhance memory and learning. Testing can also positively affect learning indirectly. Indirect effects of testing refer to the influences frequent testing can have on the study habits of students (Larsen, Butler, & Roediger, 2008; Roediger & Karpicke, 2006b). Frequent testing can lead to an increase in the amount of time one studies, more efficient study strategies, and the ability for students to learn from testing (especially when feedback is provided) (e.g., Leeming, 2002; Szpunar, McDermott, & Roediger, 2008). For example, if a course involves frequent testing compared to only minimal testing (e.g., a midterm and final exam), students are more likely to study, space out studying, and keep up with readings, all of which have been shown to improve memory performance and learning (Larsen et al., 2008; Roediger et al., 2010). Despite the importance of the indirect benefits of testing, this dissertation focused on the direct benefit of testing (i.e., the testing effect).

Even though the testing effect is not commonly known by those outside of cognitive psychology, the idea of enhanced retention from retrieval is not new. Bacon (1620/2000) and James (1890) both argued that active recitation or retrieval through testing was a more effective strategy for learning than simply restudying the material. Additionally, the testing effect has been studied in psychology or education for at least 100 years. Although early studies related to the testing effect were conducted with experimental techniques that may not meet today's standards, the finding that recitation (e.g., through testing) has positive effects on memory and learning has been replicated throughout the past century. Furthermore, the testing effect has been found with a variety of experimental conditions and stimuli.

The general procedure in modern research that has been used to study the testing

effect consists of three phases: (1) an initial encoding phase when participants are exposed to the stimuli, (2) a practice phase when participants are either tested on the material or restudy the material, and (3) a final test. Researchers have varied aspects of these three phases to thoroughly investigate the testing effect, such as the sequence of the study and test trials during the practice phase, the timing of the final test, the match between the practice and final test questions, and the type of stimuli. It should be noted that when discussing this general procedure, the abbreviations of "S" and "T" are commonly used to refer to a study or a test phase, respectively. Typically, the first "S" refers to the initial study or encoding phase, and the final test is not written into the abbreviations. For example, SST refers to initial encoding and a practice phase consisting of a study session and test, which would be followed by a final test.

It is common practice in learning and memory experiments to use a study-test multitrial paradigm (i.e., alternating study and test trials). Based on work by Tulving (1967), researchers have compared this alternating study and test trials condition (e.g., STST) to conditions with an emphasis on the study trials (e.g., SSSS) and on the test trials (e.g., STTT) to examine the benefit of testing on retention. The standard assumption leads to the predictions that emphasizing studying (e.g., SSSS) should enhance retention on a final test due to the additional study trials, while emphasizing testing (e.g., STTT) should decrease retention (Karpicke & Roediger, 2007). Research investigating these predictions has found that the pattern of results is dependent upon when the final test is given; specifically, whether the final test occurs immediately following the practice phase or after a delay. When the final test occurs immediately, additional study trials (e.g., SSSS) typically lead to greater retention than additional test trials (e.g., STTT) or no

difference is found between the conditions. However, contrary to the predictions based on the standard assumption, the opposite pattern has been shown for long-term retention. Instead, additional test trials not only enhance long-term retention, but testing leads to greater long-term retention than additional studying (i.e., the testing effect) (Hogan & Kintsch, 1971; Karpicke & Roediger, 2007; Thompson, Wenger, & Bartling, 1978; Wheeler, Ewers, & Buonanno, 2003).

Recent research (e.g., Butler, 2010; McDaniel, Anderson, Derbrish, & Morrisette, 2007; Rohrer, Taylor, & Sholar, 2010) has investigated whether the testing effect would still occur if the questions on the final test and those on the test(s) during the practice phase were not identical (i.e., the questions on the final test would require transfer). Transfer refers to carrying over what is learned in one context (e.g,, on the practice tests) to another context (e.g., the final test). These studies have shown that when a final test requires transfer (i.e., the questions on the final test are not identical to the questions on the test(s) during practice), the testing effect still occurs, and can even be slightly larger than when transfer is not required on a final test (Butler, 2010; McDaniel et al., 2007; Rohrer et al., 2010). These findings that the testing effect can occur for tested as well as related, untested, material is important because it demonstrates that the testing effect represents more than the mere reproduction of previous test answers.

Researchers have also used various types of stimuli to study the testing effect. The majority of the research examining the testing effect has been conducted using word and picture lists. Generally, these studies have demonstrated that testing leads to increased performance on a final test compared to restudying or not practicing the material with both types of materials (e.g., Hogan & Kintsch, 1971; Thompson et al., 1978; Wheeler et

al., 2003; Wheeler & Roediger, 1992). Additionally, the testing effect also has been found with paired-associates (i.e., learning the pairing of two items, such as non-word-word pairings like ZOF-college). Research using paired-associates has shown that testing promotes greater long-term retention than additional studying and that repeated testing leads to even greater benefits of testing (Allen, Mahler, & Estes, 1969; Carrier & Pashler, 1992; Estes, 1960; Izawa, 1970). Furthermore, a small number of studies have used educationally relevant stimuli when investigating the testing effect. The type of educationally relevant stimuli used in these studies have included foreign-language vocabulary words (Carrier & Pashler, 1992; Pashler, Zarrow, & Triplett, 2003), general knowledge questions (McDaniel & Fisher, 1991), prose materials (Roediger & Karpicke, 2006a), and video lectures (Butler & Roediger, 2007). Enhanced long-term retention due to testing was observed with these educational materials. Together, these studies demonstrate that the testing effect is a robust finding that occurs with a variety of stimuli, including educationally relevant stimuli.

While all of the evidence for the testing effect discussed thus far has been conducted in the laboratory, the testing effect has also been demonstrated in actual classrooms. Similar to the laboratory studies, classroom studies have shown the same positive effects of testing on long-term retention (Gates, 1917; Leeming, 2002; McDaniel et al., 2007; McDaniel, McDermott, Agarwal, & Roediger, 2008; Roediger, McDaniel, McDermott, & Agarawl, 2007). Classroom studies have also shown the testing effect when the tests (or quizzes) are given in the classroom (Leeming, 2002; McDaniel et al., 2007; Roediger et al., 2007) as well as online (Sun & McDaniel, 2008; McDaniel et al., 2008). Additionally, in both laboratories and classrooms, the testing effect has been

demonstrated with a variety of age groups ranging from preschoolers (e.g., Fritz, Morris, Nolan, & Singleton, 2007) to older adults (e.g., Logan & Balota, 2008). The evidence discussed above clearly demonstrates that, under the right conditions, testing can have positive effects on learning for people across the lifespan in both the laboratory and actual classrooms.

There are three important factors that can influence the testing effect, which are the format of the test, answer feedback, and the schedule for testing. In terms of the format of the test, there are two important factors that have been investigated: (1) whether the format of the final test has to be the same format as the practice test(s), and (2) whether different practice test formats vary in the degree of enhancement of long-term retention. To answer these questions, researchers have compared two types of memory tests: recognition (i.e., tests that require the identification of the correct response among the options presented) and production tests (i.e., tests that require retrieval or production of the correct response). Regardless of the format of the final test, the testing effect is seen for both recognition and production tests (Carpenter & DeLosh, 2006; Darley & Murdock, 1971; Kang, McDermott, & Roediger, 2007; Mandler & Rabinowitz, 1981). However, the magnitude of the testing effect does differ based on the format of the practice test(s). Both laboratory (e.g., Carpenter & DeLosh, 2006; Glover, 1989; Kang et al., 2007) and classroom studies (McDaniel et al., 2007) have shown that production tests lead to greater long-term retention than recognition tests. However, there is an important caveat to this finding. The memory benefits of testing, with both recognition and production tests, are contingent upon successful retrieval during the practice test, and test performance is usually much lower for production tests compared to recognition tests

(Roediger et al., 2010; Roediger & Karpicke, 2006b). Therefore, unless corrective feedback is provided, the fact that performance on production tests is typically much lower than on recognition tests causes practice recognition tests to produce greater benefits to long-term retention (Roediger & Karpicke, 2006b).

Feedback, or information about performance, is another factor that can influence the testing effect. Research has demonstrated that the testing effect occurs regardless of feedback; however, providing feedback can increase the magnitude of the testing effect (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler, Karpicke, & Roediger, 2007; Butler & Roediger, 2008; Kang et al., 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005). Generally, it has been shown that feedback can enhance the testing effect by increasing retention of initially correct responses (Butler, Karpicke, & Roediger, 2008) and correcting initially incorrect responses (Bangert-Drowns et al., 1991). Additionally, feedback needs to include corrective information for it to increase future performance (e.g., Pashler et al., 2005). Furthermore, it is has been shown that providing the correct response in the feedback message is more effective than indicating whether the given response is correct or incorrect (Bangert-Drowns et al., 1991; Pashler et al., 2005). Future performance is enhanced even more when feedback includes not only corrective information (i.e., the correct response) but also includes an explanation of the correct response (e.g., Moreno, 2004).

Another factor that can influence the testing effect is the schedule of testing, specifically the number of tests and spacing between tests. In terms of the number of tests, researchers have found that the magnitude of the testing effect can vary based on the number of tests used during the practice phase. Researchers have demonstrated this

by examining multiple practice tests relative to a single test or to multiple study trials, and have found that repeated testing can increase the benefit of testing on long-term retention compared to taking a single test or restudying (e.g., Wheeler & Roediger, 1992). In terms of the spacing between tests, three different types of spacing schedules of retrieval practice have been examined with the testing effect: (1) massed retrieval practice (i.e., where practice testing would occur one after another without any interruptions), (2) spaced retrieval practice (i.e., where a delayed practice test would occur followed by equally spaced subsequent tests) and (3) expanded retrieval practice (i.e., where an immediate practice test would occur followed by subsequent tests with the spacing between the subsequent tests gradually increasing with each test). The results of research investigating these three schedules in the context of the testing effect is consistent with the traditional cognitive findings regarding the spacing effect (i.e., massed versus spaced practice; for a review of the spacing effect, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). That is, distributed retrieval practice (i.e., spaced and expanded retrieval practice) produces greater long-term retention compared to massed retrieval practice (see Balota, Duchek, & Logan, 2007 for a review). Thus, the optimal schedule to increase the testing effect is repeated testing with distributed retrieval practice whether using an equally spaced or expanded schedule.

Theoretical Explanations of the Testing Effect

Researchers have attempted to understand the testing effect at a theoretical level, although the mechanisms underlying the benefit of retrieval practice on memory is not well understood. Two theories, additional exposure and overlearning, were first proposed to account for the facilitation of testing on long-term retention (for a review see

Dempster, 1996; 1997). Additional exposure argues that the testing effect is not surprising because it has already been demonstrated that studying information repeatedly (two or more times) leads to better retention than if it is only studied once. Thus, the testing effect may be the result of the additional exposure to the material that occurs during testing (Thompson et al., 1978). However, studies have demonstrated that the testing effect is the result of more than additional exposure (e.g., Karpicke & Roediger, 2007). When the amount of exposure time to the material for restudying and testing have been equated, people who are tested show better long-term retention on a final test compared to people who simply restudy the material (Karpicke & Roediger, 2007). Thus, processes other than additional exposure to the material must be responsible for the testing effect. Similar to the additional exposure explanation, the overlearning explanation of the testing effect is based on the idea that enhanced retention is the result of overlearning (i.e., continuing to practice material beyond the point of initial mastery) through practicing the material, or a portion of the material, with testing (Slamecka & Katsaiti, 1988; Thompson et al., 1978). Once again, the evidence makes the overlearning explanation inadequate for a couple of reasons. First, overlearning cannot account for the finding that testing leads to enhanced long-term retention, but for short-term retention, restudying the material typically leads to greater retention than testing (Roediger & Karpicke, 2006a; Wheeler et al., 2003). The overlearning explanation cannot account for this interaction because it predicts a main effect for both short- and long-term retention (Roediger & Karpicke, 2006b). Second, neither the overlearning nor the additional exposure explanations can account for the fact that the testing effect has been shown not only for the tested material but also for related material (e.g., Chan et al., 2006). Thus, it

is clear that the overlearning and additional exposure explanations are not adequate explanations of the testing effect.

Transfer-appropriate processing, an explanation based on the concept of transfer, may provide a better account for the testing effect. The notion of transfer-appropriate processing is that memory performance is based on how well the processes engaged in at encoding transfer to the processes needed at retrieval (Morris, Brandsford, & Franks, 1977). Based on this idea, McDaniel (2007) has argued that if encoding and retrieval processes are not congruent, test performance will serve as an index of transfer-appropriate processing more than a measure of learning. Transfer-appropriate processing appears to be able to explain the testing effect; practicing using testing engages the processes needed on a subsequent test. However, transfer-appropriate processing makes a prediction regarding test format that differs from the evidence with the testing effect. Transfer-appropriate processing predicts that performance will be greatest when the format of the final test matches that of the previous tests, but instead the evidence demonstrates that practice production tests lead to greater performance on a final test than practice recognition tests, regardless of the format of the final test (Carpenter & DeLosh, 2006; Kang et al., 2007; McDaniel et al., 2007). Thus, while transfer-appropriate processing seems to be a reasonable explanation for the testing effect, it cannot completely account for the testing effect literature.

The idea that some facet of the retrieval process is producing the testing effect is another explanation for the effect. There have been three major theories proposed about how the retrieval processes engaged in during testing lead to enhanced long-term retention, which are the notions of effortful retrieval, the elaboration of retrieval routes,

and the elaborative retrieval hypothesis. Effortful retrieval refers to the concept that deep, effortful initial retrieval influences subsequent retrieval with greater initial retrieval difficulty leading to greater retention on subsequent retrieval (Bjork, 1975; Bjork & Bjork, 1992). Researchers have shown the positive effects of effortful initial retrieval on retention (e.g., Auble & Franks, 1978; Gardiner, Craik, & Bleasdale, 1973), and the testing effect is a method for creating deep, effortful initial retrieval. Another way that retrieval can enhance subsequent retrieval is by increasing the retrieval routes that access the memory trace (Bjork, 1975; McDaniel, Kowitz, & Dunay, 1989; McDaniel & Masson, 1985). The testing effect could be the result of the retrieval routes of a memory trace being strengthened by initial retrieval. That is, retrieval practice can increase the number of retrieval routes in the memory representation, and thus strengthen the memory representation by providing multiple retrieval routes to access the memory representation (McDaniel & Masson, 1985). Recently, Carpenter (2009) proposed the elaborative retrieval hypothesis in an attempt to develop a theoretical explanation for the testing effect based on the general idea of the elaboration of retrieval routes. According to the elaborative retrieval hypothesis, retrieval practice will not only strengthen the retrieval routes of a memory trace, but importantly will also activate and strengthen the connections between the cue, target and other related information providing multiple, elaborative retrieval routes to access the target information in the future (Carpenter, 2009). Theories based on effort and elaboration (i.e., effortful retrieval, the elaboration of retrieval routes, and the elaborative retrieval hypothesis) seem to provide good accounts of the testing effect to date; however, all of these explanations are vague in terms of the specific mechanisms underlying the testing effect, and there may also be other

mechanisms responsible for or that lead to the testing effect and consequently other theoretical accounts.

More recent and related theories have been proposed to describe the mechanisms underlying the testing effect, which are the mediator effectiveness hypothesis (Pyc & Rawson, 2010), semantic mediator hypothesis (Carpenter, 2011) and mediator shift hypothesis (Pyc & Rawson, 2012). The basic notions related to all of these hypotheses is that a mediator is anything (e.g., word, phrase, concept) that links a cue to a target, and the effectiveness of mediators is influenced by two factors: mediator retrieval or the ability to retrieve the mediator, and mediator decoding or the ability of the mediator to elicit the target memory. The mediator effectiveness hypothesis (Pyc & Rawson, 2010) posits that the testing effect occurs because during practice testing people generate and use more effective mediators. The mediator shift hypothesis (Pyc & Rawson, 2012) suggests that the benefits of retrieval practice occur because when engaging in retrieval practice people modify their mediators when they experience retrieval failure during practice. These two hypotheses are thought to complement one another, and taken together, they posit that effective mediators can lead to the testing effect because retrieving the mediators during retrieval practice may strengthen the memory trace, and retrieval failure during retrieval practice can lead to a change in the mediator to a more effective mediator (Pyc &Rawson, 2010; 2012). Carpenter (2011) provided support for the mediator effectiveness hypothesis as well as demonstrated the benefit of mediator use when participants were not specifically asked to generate mediators. Based on these findings, Carpenter (2011) concluded that the semantic mediator hypothesis provides a useful theoretical account for the testing effect. The semantic mediator hypothesis

(Carpenter, 2009; 2011; Carpenter & DeLosh, 2006) is similar to the mediator effectiveness hypothesis, and it posits that the testing effect occurs because practice testing increases the likelihood that related, semantic information is activated, which can serve as a semantic mediator for later retrieval. Evidence has been found for the contribution of mediator retrieval and mediator decoding to the testing effect (e.g., Carpenter, 2011; Pyc & Rawson, 2010; 2012); however, these hypotheses acknowledge that effective mediators are not the only mechanism underlying the testing effect.

Until recently, theories based on elaboration seem to provide the best account of the testing effect, although transfer-appropriate processing and the mediator hypotheses provide insight into the mechanisms underlying the testing effect. However, a recent study conducted by Karpicke and Smith (2012) challenge the notion that the benefits of retrieval practice are simply due to elaboration. Karpicke and Smith (2012) conducted a series of experiments investigating whether the testing effect can be attributed to elaborative encoding by comparing retrieval practice to various elaborative study conditions. In all of their experiments, they found that retrieval practice lead to superior memory on a final test compared to the elaborative study conditions, even when the stimuli used reduced or prohibited the production of elaborations/mediators with retrieval practice (Karpicke & Smith, 2012). Based on these results, Karpicke and Smith (2012) concluded that the benefit of retrieval practice can best be explained by retrieval-specific mechanisms rather than elaborative mechanisms. Specifically, Karpicke and Smith (2012) argue that their findings support the notion that retrieval practice benefits memory through the improvement of the diagnostic value of retrieval cues (Karpicke & Blunt, 2011; Karpicke & Zaromb, 2010). The idea behind this cue diagnosticity perspective is

that retrieval is a decision discrimination process where successful retrieval is a function of how effectively a retrieval cue can specify a particular candidate (i.e., the target) while excluding other potential candidates (Moscovitch & Craik, 1976; Nairne, 2002; Tulving, 1974; Tulving & Thomson, 1973). This notion of cue diagnosticity can be applied to the testing effect in that retrieval practice could be enhancing the diagnostic value of retrieval cues (e.g., through the restriction of the set of candidates to be included in the search set, by enhancing how well a retrieval cue specifies a particular candidate, etc.) as opposed to an increase or addition of the number of encoded features as occurs with elaboration (Karpicke & Blunt, 2011; Karpicke & Smith, 2012; Karpicke & Zaromb, 2010). However, the goal of this dissertation was not to directly test or compare these particular theories. Instead, this dissertation explored whether the dual-process signal detection model (Yonelinas, 1994) can be used to examine the phenomena related to the testing effect, and consequently possibly provide a useful technique for continuing to tease apart these theories and even possibly formulate new theories to explain how and why retrieval practice benefits long-term memory.

Dual-Process Signal Detection Model

The notion of dual process theory is that recognition memory is based on two different processes or types of memory: recollection and familiarity. Recognition memory judgments can be made on the retrieval of specific aspects (e.g., the context) of a study event (i.e., recollection), or based on a feeling of knowing, but without retrieval of specific qualitative information about the study event (i.e., familiarity). For example, if you see someone you know and realize that you know him/her but cannot recall his/her name, that would be familiarity; recollection is being able to retrieve contextual

information from memory such as the person's name or where you met the person. Recognition memory judgments are thought to always involve familiarity, whereas only some recognition memory judgments will involve recollection or only some items will be recollected (Parks & Yonelinas, 2008; Yonelinas, 2002). Several models (see Yonelinas, 2002 for a review) have been proposed that make different predictions about the functional nature and neural substrates underlying these two processes, and consequently how these processes are measured (Yonelinas, 2001a; 2002). One of the most prominent models, the dual-process signal detection model (Yonelinas, 1994), describes recollection and familiarity in terms of response confidence. This model was the focus of this dissertation.

The dual-process signal detection (DPSD) model (Yonelinas, 1994) of recognition memory is a hybrid model, meaning that it integrates signal detection theory and threshold theory. The DSPD model makes four assumptions regarding recollection and familiarity. The first assumption is that familiarity reflects a signal-detection process such that it is always thought to be successful in some way (i.e., there is always a memory signal that provides some useful information even if it doesn't lead to an accurate response); the familiarity distributions for old and new items, which are both normal (or Gaussian) in shape, are overlapping due to variability in memory strength (Parks & Yonelinas, 2008; Yonelinas, 2001a). Support for this assumption has come from studies examining recognition performance under conditions where performance should rely exclusively on familiarity, such as in amnesics who are unlikely to be able to recollect specific details about an event (i.e., they cannot rely on recollection) but can make memory judgments based on familiarity (e.g., Yonelinas, Kroll, Dobbins, Lazzara, &

Knight, 1998).

In contrast, the second assumption is that recollection is a threshold process, meaning that it either occurs or fails; that is, qualitative information about the study event will either be retrieved or will not be retrieved (Yonelinas, 1994; 2001a). This assumption has been investigated using experimental conditions where familiarity can play a limited role in performance (e.g., tests of associative recognition and source memory) and examining the shapes of receiver operating characteristics, typically abbreviated as ROCs. ROCs are graphical functions that relate the proportion of correctly recognized items to the proportion of incorrectly recognized items and can indicate variations in response bias. These studies (e.g., Kelley & Wixted, 2001; Yonelinas, 1997; 1999) have supported the notion that recollection reflects a threshold process by demonstrating that the ROCs produced in these conditions are relatively linear compared to the curvilinear ROCs that are observed with tests of recognition memory that rely primarily on familiarity or both recollection and familiarity, such as item recognition (Parks & Yonelinas, 2008; Yonelinas, 2001a; 2001b). Furthermore, this finding demonstrates that recognition performance that relies primarily on recollection cannot be accounted for using signal-detection processes because curvilinear ROCs (i.e., not linear ROCs) are always predicted with signal detection theory (Yonelinas, 2001a).

A third assumption of the DPSD model is that recollection leads to high confidence responses, whereas familiarity can lead to a wider range of confidence responses (Yonelinas, 2002). This assumption is based on the idea that if someone can retrieve qualitative information about a studied event then they should be confident that the event occurred, but people may be less confident about familiarity-based memory

judgments because of the overlap of the familiarity distributions for old and new items (Yonelinas, 2001a; 2001b).

The final assumption is that recollection and familiarity are two independent processes (Parks & Yonelinas, 2008; Yonelinas, 1994). This assumption has been supported by numerous behavioral studies (for reviews, see Jacoby, Yonelinas, & Jennings, 1997; Yonelinas, 2002). For example, behavioral studies have shown that different behavioral manipulations can affect recollection but not familiarity, or vice versa. For instance, divided attention and list length manipulations have been found to disproportionately influence recollection but not familiarity (e.g., Jacoby, 1991), while manipulations of response bias and study-test lag have been found to disproportionately affect familiarity (e.g., Yonelinas & Levy, 2002). Additionally, results from studies conducted using event related potentials (ERPs), functional magnetic resonance imaging (fMRI), and brain-damaged patients have also provided support for the notion that recollection and familiarity reflect independent processes (see Yonelinas, 2002 for a review).

Theoretical questions about the nature of recollection and familiarity have been addressed using various measurement methods, such as the process dissociation procedure (Jacoby, 1991), remember-know procedure (Tulving, 1985), and receiver operating characteristics procedure (Yonelinas, 1994). The receiver operating characteristics (ROCs) procedure (Yonelinas & Parks, 2007) is one of the most direct ways to estimate the recollection and familiarity processes (Yonelinas, 2002; Yonelinas & Parks, 2007), and this type of assessment was a prominent analysis for the experiments conducted in this dissertation. As described earlier, an ROC is a function that relates the

hit rate (i.e., the proportion of correctly recognized target items) to the false alarm rate (i.e., the proportion of incorrectly recognized lure items) across response bias (Macmillan & Creelman, 2005). ROCs are derived using multiple points that are collected under different levels of response bias (Macmillan & Creelman, 2005). There are several ways to obtain the multiple points in ROCs with the most common method being the confidence rating method where participants are required to rate the confidence of their recognition judgments (Parks &Yonelinas, 2008; Yonelinas & Parks, 2007). With the confidence rating method, ROCs are plotted as a function of response confidence with the leftmost point reflecting the most confidently recognized items and recognition confidence decreasing for each subsequent point.

Performance can be assessed by examining the shape of the ROC. The greater the area under the curve (i.e., the more the function is towards the upper left corner) the greater the memory discriminability (i.e., performance), whereas chance performance (i.e., when the hit rate is equal to the false alarm rate) is reflected by a function lying on the diagonal. The shape of the ROC can be quantified by examining the ROC in z-space or by plotting the z-score of each hit and false alarm rate to produce a zROC. If the zROC is linear, then the y-intercept can be used as a rough index of memory discriminability and the slope can be used as an index of the symmetry of the ROC (Parks & Yonelinas, 2008; Yonelinas & Parks, 2007). It has been shown that the shape of the ROC can also reflect the contribution of recollection and familiarity with recollection and familiarity producing distinct ROCs; when performance is above chance, the ROCs produced under conditions relying primarily on familiarity are typically curvilinear and symmetrical, whereas the ROCs produced under conditions relying primarily on recollection are

typically more linear and asymmetrical (Yonelinas, 2002). Theoretically based models (e.g., the DPSD model) can be fitted to the data to obtain estimates of recollection and familiarity (Yonelinas & Parks, 2007). With the DPSD model, the probability of making a "yes" response is described by the following equations:

$$(1)\ P(\text{"yes"} \mid \text{studied})_i = R + (1 - R)\Phi(d' - c_i)$$

and

$$(2)\ P(\text{"yes"} \mid \text{new})_i = \Phi(-d' - c_i)$$

where R refers to the recollection parameter, *d'* refers to the familiarity parameter, c_i refers to the response criterion, and Φ refers to the normal cumulative distribution function which signifies the proportion of the target and lure distributions that exceed the response criterion (c_i) given that the distance between the means of the two Gaussian distributions is *d'* (Macmillan & Creelman, 2005; Yonelinas, 1998). It should be noted that since the DPSD model describes recollection as a threshold process and familiarity as a signal detection process, recollection is calculated as a probability whereas familiarity is calculated in *d'* units in these equations (Macmillan & Creelman, 2005; Yonelinas & Parks, 2007). Using these theoretically-based equations, the DPSD model can be directly fitted to the observed data to obtain parameter estimates of recollection and familiarity. The ROCs procedure has been effectively applied to various recognition memory paradigms and has been useful in helping researchers understand the memory processes underlying recognition (see Macmillan & Creelman, 2005 for ROC details and Yonelinas & Parks, 2007 for DPSD details).

The DPSD model and the analysis of ROCs may provide a solid theoretical framework and method for examining the processes of familiarity and recollection in

relation to the testing effect. A study by Chan and McDermott (2007) examined whether practice testing can influence the underlying memory processes of recognition judgments even when it does not impact recognition hit rates. They investigated this using a testing effect paradigm where studied information was either practiced (using free recall) or not practiced, and performance on an immediate final test was examined. As a reminder, on an immediate final test, there is typically no benefit of practice testing reflected in the hit rates (i.e., the benefit of practice testing on performance is not typically seen with short-term retention). Across three experiments using various measurement procedures (i.e., the remember/know and process dissociation procedures), Chan and McDermott found that practice testing changed how recognition judgments were made even when there was no difference in recognition hit rates. Specifically, they found that practice testing enhanced the probability of later recognition by recollection (but did not influence the contribution of familiarity) independent of enhancement to the recognition hit rates (Chan & McDermott, 2007).

Recently, Verkoeijen, Tabbers, and Verhage (2011) conducted a follow-up study to Chan and McDermott's (2007) study to examine how practice testing affects the processes of recollection and familiarity in comparison to restudying since practice testing was only compared to no practice by Chan and McDermott (2007). The procedure used by Verkoeijen et al. (2011) was very similar to the third experiment of Chan and McDermott's (2007) study using the process dissociation procedure, although recollection and familiarity were estimated using both the process dissociation procedure (Jacoby, 1991) and extended measurement model (Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995). More specifically, Verkoeijen et al. (2011) examined recollection and

familiarity using the process dissociation procedure when information was practiced using testing or restudying (manipulated both within- and between-subjects), feedback was given on the practice test, and stronger cues were used on the practice test. It should be noted that similar to Chan and McDermott (2007), an immediate final test was used and thus in all of their experiments recollection and familiarity were examined when testing had no benefit to performance. In line with the findings from Chan and McDermott (2007), Verkoeijen et al. (2011) found that practice testing increased recollection but not familiarity; however, practice testing only increased recollection to a greater extent than restudying when feedback was given on the practice test or stronger cues were used on the practice test. Furthermore, they found that restudying increased the contribution of familiarity to recognition judgments compared to practice testing in all four experiments (Verkoeijen et al., 2011). While both of these studies (Chan & McDermott, 2007; Verkoeijen et al., 2011) found that practice testing increases recollection in comparison to no practice and even restudying, under the right circumstances, it is unclear whether similar results would be found when practice testing does influence recognition performance (i.e., with long-term retention). Furthermore, it would be useful to examine how practice testing influences recollection and familiarity using a more direct measurement method (i.e., the receiver operating characteristics procedure).

CHAPTER 3

OVERVIEW OF EXPERIMENTS

Overall, this dissertation investigated whether the DPSD model (Yonelinas, 1994) could provide a useful method for investigating the testing effect. Specifically, this study addressed three questions: (1) Can the DPSD model be used to investigate the testing effect (i.e., the direct benefit of retrieval practice on long-term retention)? (2) Will the number of practice sessions affect the parameter estimates of recollection and familiarity? (3) Does practice testing increase the contribution of recollection and/or familiarity based on what is needed on the final test? These questions were investigated across two experiments.

Experiment 1 investigated the first two questions of whether the DPSD model can be used to examine the testing effect and if the number of practice sessions would influence the parameter estimates of recollection and familiarity using a 2 (practice condition: restudying versus testing) x 2 (number of practice sessions: one versus three) between-subjects design. Experiment 2 addressed the third question of whether practice testing increases the contribution of recollection and/or familiarity based on the format of the final test using a 2 (practice condition: restudying versus testing) x 2 (final test format: standard versus recollection-dependent) between-subjects design. The dependent variables for both experiments were the proportion of correct responses, *d'* values, ROCs, and parameter estimates of recollection and familiarity. All of these dependent variables were calculated based on the confidence rating data from the final test.

CHAPTER 4

EXPERIMENT 1

Experiment 1 investigated the utility of the DPSD model of recognition memory for examining the testing effect using the ROCs procedure to obtain parameter estimates of recollection and familiarity. To be a useful method for investigating the testing effect, the DPSD model should be able to account for not only the basic testing effect but also manipulations that influence the testing effect, such as the number of practice sessions. Therefore, this experiment also investigated whether the number of practice sessions would affect the parameter estimates of recollection and/or familiarity. An increase in the number of practice tests increases the magnitude of the testing effect (e.g., Wheeler & Roediger, 1992). Furthermore, increasing study duration by increasing the duration of each study item or repeating items (as is the case with multiple practice sessions) has been shown to lead to comparable increases in recollection and familiarity, with slightly larger increases in recollection, with various measurement methods (see Yonelinas, 2002). To investigate these questions, participants studied a list of words, completed a brief distractor task, practiced the material either once or three times using either restudying or testing, and then took a delayed final old/new test. On the final old/new test, participants made confidence rating responses using a scale from 1 to 6 with 1 being "sure it's a new word" and 6 being "sure it's an old word". Confidence ratings from the final old/new test were used to measure the proportion of correct responses, calculate *d'* values, create ROC and zROC curves, and estimate recollection and familiarity, all of which were the dependent variables.

If the DPSD model offers a method for examining the testing effect, then the

following pattern of results would be observed. The enhancement of long-term retention with practice testing should lead to an increase in the contribution of recollection, familiarity or both to recognition memory judgments, and depending on whether the contribution of recollection, familiarity or both processes is increased, the shape of the ROCs should reflect this and be in line with the predictions of the DPSD model (i.e., the shape of the ROCs the DPSD model would predict based on whether recollection, familiarity or both processes is increased). Additionally, for the DPSD model to be a useful method for investigating the testing effect, it should also be able to account for the enhancement of the benefit of practice testing with multiple practice tests. Thus, in addition to the general pattern of results just described, the parameter estimates of recollection and/or familiarity should increase for practice testing compared to restudying with an increased contribution of the process(es) (i.e., recollection, familiarity or both) for three practice sessions while maintaining the predictions of the DPSD model. However, if the DPSD model does not provide a useful technique for investigating the testing effect, then there are a few patterns of results that could occur. First, practice testing for long-term retention could increase the contribution of recollection, familiarity or both, but the ROCs will not reflect this as the DPSD model would predict. Second, practice testing and restudying could influence the parameter estimates of recollection and familiarity in the same way. Third, if the DPSD model cannot account for the enhancement of the testing effect with multiple practice tests then there would be no difference in the parameter estimates between multiple practice tests and a single practice test. If any of these possibilities occur, then the DPSD model would not offer a useful method for investigating the testing effect. In spite of these possibilities, it was

hypothesized that the results of Experiment 1 would be in line with the DPSD model (i.e., an increased contribution of recollection, familiarity or both would be found for practice testing compared to restudying with this enhancement being more pronounced with multiple practice tests, and the ROCs would reflect the patterns that the DPSD model would predict), and thus provide evidence that the DPSD model can be used to investigate the testing effect.

Method

Participants. One hundred (62 females, 38 males with a mean age of 20.8) participants from the student population of the University of Nevada, Las Vegas via the Department of Psychology's subject pool were recruited for this experiment. For their participation, students were compensated with credit that could be applied to a psychology course. The only restrictions for participation were that one needed to be at least 18 years of age at the time of participation and able to fluently speak and understand English.

Materials. The stimuli used in Experiment 1 were a list of 240 low frequency (1-40) words taken from the MRC Psycholinguistic Database (Kucera & Francis, 1967). Low frequency words were used because they (relative to high frequency words) have been shown to lead to greater hit rates and lower false alarm rates (e.g., Glanzer & Adams, 1985). All of the words were 4 to 6 letters in length, 1 to 2 syllables, and concrete nouns. The word list was randomly separated into two lists, List 1 and 2, containing 120 words each. List 1 was used for the study items during the encoding phase, to create the two-letter stems for the practice test, and for the old items on the final old/new test; List 2 was used for the new items on the final old/ new test.

Procedure. Experiment 1 consisted of four phases: encoding, distractor task, practice, and final old/new test. Before beginning, participants were informed that they would be asked to memorize a list of words and later be tested on them. In the encoding phase, participants were presented with the words from List 1 in a random order, with each word being presented one at a time for 4 seconds. After being presented with the entire list, participants completed a brief distractor task (which took approximately 2 minutes) consisting of counting the number of 'X's on the screen. Following the distractor task, participants practiced the material either once or three times using either restudying (i.e., either SS or SSSS) or testing (i.e., either ST or STTT). The type of practice (restudying versus testing) and number of practice sessions (one versus three) were counterbalanced across participants. It should be noted that the type of practice (i.e., restudying or testing) was used for all practice sessions. Also, following each practice, participants in both practice conditions completed another distractor task (i.e., counting the number of 'X's on the screen for approximately 2 minutes). In the restudying condition, participants were presented with List 1 again in the same manner as during the encoding phase (i.e., one word at a time for 4 seconds), and in the testing condition, participants were tested on the words from the encoding phase. The practice test consisted of two-letter stems for the 120 studied words from List 1 (e.g., 'is_ _ _ _' for the word island), and the stems were presented one at a time, remaining on the screen until the participant responded. Regardless of the accuracy of the participant's response, feedback was presented immediately after each response; feedback indicated the correct studied word that completed the stem (e.g., "The correct response is island"), and was displayed for 500 ms. The presentation order of the words was randomized for both

practice conditions during all practice sessions.

Finally, after a two-day delay, participants completed a final old/new test. The final old/new test consisted of all 240 words (i.e., Lists 1 and 2), which were mixed together and presented one word at a time in a randomized order. Additionally, each word was preceded and followed by a white and red fixation cross, respectively. The white fixation cross was presented for 500 ms, followed by the word for 1,500 ms, and the red fixation cross remained on the screen until participants made their response. Participants responded that the word was old or new using a scale from 1 to 6, where 1 corresponded to 'sure it's a new word', 2 corresponded to 'somewhat sure it's a new word', 3 corresponded to 'guessing but think it's a new word', 4 corresponded to 'guessing but think it's an old word', 5 corresponded to 'somewhat sure it's an old word', and 6 corresponded to 'sure it's an old word'. Participants made their response on a standard computer keyboard using the numbers 1 through 6.

Analyses

The dependent measures were the proportion of correct responses, *d'* values, ROCs, and parameter estimates of recollection and familiarity, which were calculated using the confidence rating data from the final old/new test. The practice test data were not included in any of the analyses; however, the proportion of correct responses from the practice tests is summarized in the Appendix. Performance on the final old/new test was examined using both proportion of correct responses and *d'* values. The typical measure of performance used in the testing effect literature is accuracy calculated as the proportion of correct responses. To obtain the proportion of correct responses for each participant, responses of 4, 5 and 6 were counted as an 'old' response and responses of 1,

2 and 3 were counted as a 'new' response. The most widely used performance measure of detection theory is *d′*, which offers a performance measure that is roughly invariant to manipulations of response bias unlike measuring performance as the proportion of correct responses (Macmillan & Creelman, 2005). For this study, the measure of sensitivity used was the *d′* meaure of signal-detection theory that assumes underlying unequal-variance distributions, d_a', which is calculated in terms of z (i.e., the inverse of the normal distribution function) as

(3) $d_a' = (2/1 + s^2)^{1/2}[z(H) - sz(F)]$

where *z*(H) is the z score of the hit rate, *z*(F) is the z score of the false alarm rate, and s is the slope of the zROC (Macmillan & Creelman, 2005). For this study, the hit rate reflected the proportion of studied items accepted as old, and the false alarm rate reflected the proportion of new items accepted as old.

In addition to calculating proportion of correct responses and *d'*, ROCs were produced for each condition by plotting the average hit rate against the average false alarm rate as a function of response confidence (Yonelinas & Parks, 2007). Following the ROCs procedure, the ROCs were plotted such that the leftmost point reflected the most confidently remembered items (i.e., items eliciting a '6' response) and recognition confidence decreased for each subsequent point. That is, the 6-point confidence scale used on the final old/new test yielded a 5-point ROC with the leftmost point reflecting the most confidently remembered items (i.e., hits = P[6 | old], false alarms = P[6 | new]), and each subsequent point was calculated by including the next most confidently recognized items:

(4a) hits = P[6 | old] + P[5 | old]

(4b) false alarms = P[6 | new] + P[5 | new]

(5a) hits = P[6 | old] + P[5 | old] + P[4 | old]

(5b) false alarms = P[6 | new] + P[5 | new] + P[4 | new]

(6a) hits = P[6 | old] + P[5 | old] + P[4 | old] + P[3 | old]

(6b) false alarms = P[6 | new] + P[5 | new] + P[4 | new] + P[3 | new]

(7a) hits = P[6 | old] + P[5 | old] + P[4 | old] + P[3 | old] + P[2 | old]

(7b) false alarms = P[6 | new] + P[5 | new] + P[4 | new] + P[3 | new] + P[2 | new].

The shapes of the ROCs were quantified by plotting the ROCs in z-space (i.e., plotting the z-score of each hit and false alarm rate to produce a zROC) to examine memory discriminability and the symmetry of the ROC. The intercept of the zROC provides a rough index of discriminability (*d'*), and the slope of the zROC reflects the symmetry of the ROC where a perfectly symmetrical ROC will have a slope of 1.0 and an asymmetrical ROC will have a slope either greater or less than 1.0 (Macmillan & Creelman, 2005).

Finally, the DPSD model was used to fit the observed confidence rating data for each participant to obtain parameter estimates of familiarity and recollection. Using the theoretically based equations of the DPSD model mentioned earlier,

(1) $P(\text{"yes"} \mid \text{studied})_i = R + (1 - R)\Phi(d' - c_i)$

and

(2) $P(\text{"yes"} \mid \text{new})_i = \Phi(-d' - c_i)$

the DPSD model was fitted to the data. Specifically, for each condition, these equations were fitted to the observed ROCs (i.e., the 5 points on the ROCs) using a log-likelihood estimation method (see Parks, Murray, Elfman & Yonelinas, 2011 for a recent study

using this method). The log-likelihood estimation method fits the data with the DPSD model by maximizing the log-likelihood value between the predicted function and the observed data while varying the recollection parameter, familiarity parameter, and response criterion. For this study, the solver function in Microsoft Excel was used to find the best fitting parameters (i.e., parameter estimates) for these equations by maximizing the log-likelihood value between the predicted and observed data. This was done for each participant's data to obtain parameter estimates of recollection and familiarity when the equations for the DPSD model were fitted to the observed confidence rating data. The group average of the parameter estimates of recollection and familiarity were what was analyzed using an analysis of variance (ANOVA) to determine how the processes of recollection and familiarity were affected by the experimental manipulations (i.e., for Experiment 1, restudying versus practice testing and one versus three practice sessions).

Results and Discussion

Proportion of correct responses, *d'* values and the parameter estimates of familiarity and recollection from the final old/new test were each analyzed using a 2 (practice condition: restudying versus testing) x 2 (number of practice sessions: one versus three) between-subjects ANOVA. For all analyses, an alpha level of .05 was used to determine statistical significance. The results for both the proportion of correct responses data and *d'* values demonstrated the patterns seen in the testing effect literature (see Figures 1 and 2). That is, practice testing significantly enhanced performance (i.e., higher proportion of correct responses and *d'* values) on the final old/ new test compared to restudying, regardless of the number of practice sessions, $F(1, 99) = 23.732$, $p < .001$, $\eta_p^2 = .198$ and $F(1, 99) = 25.321$, $p < .001$, $\eta_p^2 = .209$, respectively for proportion of

correct responses and *d'*. Additionally, a significant main effect of number of practice sessions was found demonstrating enhanced long-term retention (i.e., significantly higher proportion of correct responses and *d'* values) for three practice sessions compared to one practice session, regardless of the type of practice, $F(1, 99) = 15.078$, $p < .001$, $\eta_p^2 = .136$ and $F(1, 99) = 19.445$, $p < .001$, $\eta_p^2 = .168$, respectively for proportion of correct responses and *d'*. Finally, a significant interaction between practice condition and number of practice sessions was found for *d*' values, $F(1, 99) = 6.338$, $p = .013$, $\eta_p^2 = .062$; however, the interaction did not reach significance for the proportion of correct responses data, $F(1, 99) = 2.660$, $p = .106$, $\eta_p^2 = .027$. Planned comparisons for both accuracy measures were conducted using independent-samples *t*-tests. Enhanced performance with both proportion of correct responses and *d'* values was found for practice testing compared to restudying with both one practice session, $t(48) = 2.776$, $p = .008$ and $t(48) = 2.601$, $p = .012$, and three practice sessions, $t(48) = 4.004$, $p < .001$ and $t(48) = 4.312$, $p < .001$. Importantly, significantly higher proportion of correct responses and *d'* values was found for three practice tests compared to a single practice test, $t(48) = 3.664$, $p = .001$ and $t(48) = 4.392$, $p < .001$, while no significant difference in performance (i.e., proportion of correct responses or *d'* values) was found between a single restudying session versus three restudying sessions, $t(48) = 1.710$, $p = .094$ and $t(48) = 1.539$, $p = .130$. These findings replicate previous research (e.g., Wheeler & Roediger, 1992; for reviews, see Roediger & Butler, 2011; Roediger & Karpicke, 2006b) demonstrating that practice testing enhances later memory performance compared to restudying, with this benefit being further enhanced with multiple practice sessions. Furthermore, these performance results highlight not only the benefit of practice testing to long-term

memory but also demonstrate that repeated practice (e.g., with three practice sessions) only seems to truly be beneficial to long-term memory when an effective mnemonic technique (such as practice testing) is used to practice the information.

The type of practice and the number of practice sessions led to differences in the parameter estimates of recollection and familiarity (calculated from modeling the confidence rating data using the DPSD model), with these differences being reflected in the ROCs and zROCs in a manner consistent with the DPSD model (see Figures 3-6). Specifically, the analysis of the familiarity parameter estimates (see Figure 3) showed a significant increase in the contribution of familiarity for practice testing compared to restudying, regardless of the number of practice sessions, $F(1, 99) = 17.691$, $p < .001$, $\eta_p^2 = .156$. Additionally, a significant increase in the contribution of familiarity was found with three practice sessions compared to one practice session, regardless of the type of practice, $F(1, 99) = 7.588$, $p = .007$, $\eta_p^2 = .073$. While no interaction between practice condition and number of practice sessions was found, $F(1, 99) = 1.267$, $p = .263$, $\eta_p^2 = .013$, planned comparisons conducted using independent-samples *t*-tests showed that three practice sessions (compared to one practice session) significantly increased the contribution of familiarity for the testing group, $t(48) = 2.199$, $p = .033$, but not for the restudying group, $t(48) = 1.731$, $p = .090$. Additionally, practice testing (compared to restudying) significantly increased the contribution of familiarity for both a single practice session, $t(48) = 3.614$, $p = .001$, and three practice sessions, $t(48) = 2.947$, $p = .005$.

The analysis of the recollection parameter estimates (see Figure 4) showed a significant increase in the contribution of recollection for three practice sessions

compared to one practice session, regardless of the type of practice, $F(1, 99) = 4.572$, $p = .035$, $\eta_p^2 = .045$. Although only a marginally significant main effect of practice condition was found, $F(1, 99) = 3.827$, $p = .053$, $\eta_p^2 = .038$, a significant interaction between practice condition and number of practice sessions was found, $F(1, 99) = 4.538$, $p = .036$, $\eta_p^2 = .045$. The planned comparisons conducted using independent-samples *t*-tests showed that three practice sessions (compared to one practice session) significantly increased the contribution of recollection for the testing group, $t(48) = 2.700$, $p = .010$, but not for the restudying group, $t(48) = 0.006$, $p = .995$. Furthermore, a significant difference in the contribution of recollection between practice testing and restudying was found with three practice sessions, $t(48) = 2.535$, $p = .015$, but not with a single practice session, $t(48) = 0.147$, $p = .884$. Taken together, the results from the parameter estimates suggest that enhanced performance with practice testing generally leads to an increase in the contribution of familiarity; however, when multiple practice sessions are used, practice testing also increases the contribution of recollection.

Generally, the increases in familiarity and recollection (based on both the practice condition and number of practice sessions) just described are reflected in the ROCs and zROCs for each condition in a manner that is consistent with what the DPSD model would predict (see Figures 5 and 6). There were four patterns in the shapes of the ROCs and zROCs that were in line with what the DPSD model would predict for tests of item recognition (Parks & Yonelinas, 2008). (1) First, overall the ROCs were curvilinear and asymmetrical along the negative diagonal reflecting the contribution of both familiarity and recollection; the zROCs were approximately linear. (2) Additionally, the increase in familiarity but not recollection for the one practice testing condition compared to the one

restudying condition was reflected as a somewhat more curvilinear and symmetrical (along the negative diagonal) ROC with a more gradual increase towards 1, 1 for the one practice testing condition. (3) The shape of the ROCs for the restudying conditions were generally similar in shape (except that the three session restudying condition was slightly shifted towards the upper left corner reflecting the slightly higher, though not significantly different, performance) because there were no significant differences in the contributions of familiarity or recollection between the restudying conditions. (4) Finally, the significant increase in both recollection and familiarity for the three session practice testing condition was reflected in the ROC being both shifted up towards the upper left corner (reflecting higher performance) and pushed up on the left side (reflecting the increased contribution of recollection) making it more asymmetrical along the negative diagonal and the zROC U-shaped.

In addition to the accuracy and parameter estimates data, the shapes of the ROCs and zROCs being consistent with what the DPSD model would predict (based on the differences in the parameter estimates between the various conditions) provide initial evidence that the DPSD model can be used to examine not only the testing effect but also manipulations that influence the testing effect (i.e., the number of practice sessions). Thus, these findings suggest that the DPSD model can be a useful method for investigating the testing effect. However, based on the results of Experiment 1, it is unclear whether (a) practice testing increases familiarity in general as well as increases both familiarity and recollection when multiple practice tests are used, or (b) if the processes of familiarity and recollection are increased by practice testing based on what is necessary for optimal performance on the final test. There are two plausible

interpretations for the results observed in Experiment 1. First, it is possible that practice testing generally leads to an increase in familiarity (i.e., the benefit to long-term recognition memory with practice testing is due an increase in familiarity in general). Thus, the results of Experiment 1 are due to the fact that practice testing increases the process of familiarity in general. However, it is also possible that practice testing increased familiarity in Experiment 1 because familiarity is sufficient for performing well on a test of item recognition. That is, to accurately identify an item as old or new on an item recognition test, one does not necessarily need to recollect specific details of the study event (i.e., use recollection), but instead can rely on the familiarity of the item to judge whether it is old or new. Based on this, it is possible that practice testing may increase familiarity and/or recollection based on the format of the final test and which process(es) are needed to perform optimally on the test. These two possibilities were further examined in Experiment 2.

CHAPTER 5

EXPERIMENT 2

The first purpose of Experiment 2 was to replicate the basic findings of Experiment 1. The second was to examine whether a testing effect would occur when using a final old/new test where performance was more dependent on recollection. The reason for this is that while Experiment 1 demonstrated that practice testing increased the contribution of familiarity, recollection was only increased when three practice tests were used; thus, it is unclear whether practice testing increases familiarity in general or if the processes of recollection and familiarity are increased by practice testing based on what is necessary for performance on the final test. On the one hand, the pattern of results seen in Experiment 1 could have occurred because the testing effect resulted in an increase in familiarity, generally, and the increase in both familiarity and recollection seen with three practice tests was simply due to the increase in study duration and the use of generation on the practice tests. Increasing study duration using distributed presentations and generation during study compared to reading have both been shown to lead to increases in both familiarity and recollection with slightly larger increases for recollection than familiarity (Yonelinas, 2002). On the other hand, it is also possible that the testing effect increases the contribution of recollection, familiarity or both based on the demands of the final test (i.e., practice testing could increase the processes of familiarity and/or recollection differently based on the format of the final test).

The goal of Experiment 2 was to investigate these two possibilities further by manipulating the type of lure items used on the final old/new test. On the standard final old/new test (i.e., the version used in Experiment 1), novel words were used for the lure

items. However, this type of test could promote the reliance on familiarity because a specific recollection of the studied event may not be necessary to determine that a word was old or new because the novel lures were semantically different from the old words. To create a recollection-dependent final old/new test, plurality-reversed lures (see Kapucu, Macmillan, & Rotello, 2010; Rotello, Macmillan, & Van Tassel, 2000) were used. Plurality-reversed refers to the idea that a lure is created by changing an old/studied word from singular to plural (e.g., studying 'frog' and being tested on 'frogs'), or from plural to singular (e.g., studying 'computers' and being tested on 'computer'). In this case, on the final test, familiarity (due to the semantic similarity between the old words and similar lures) should not be sufficient to judge whether a specific word was previously studied; instead, people should have to recollect the specific word, including its singularity/plurality. Previous dual-process research has shown that on recognition tests that include similar lures (e.g., plurality-reversed lures), the recollection and familiarity processes are supplemented by a slow, accurate process referred to as the recollect-to-reject process (e.g., Rotello et al., 2000). The basic idea behind the operation of this recollect-to-reject process is that a studied item is recalled to reject the similar foil that cannot be recalled (Rotello et al., 2000; Yonelinas, 1997). Furthermore, this recollect-to-reject process does not influence recognition judgments for novel lures (as it is not needed to reject them) as these items are not similar to any studied item nor have they been previously seen (Rotello et al., 2000; Yonelinas, 1997).

If the testing effect increases the contribution of recollection, familiarity or both based on the format of the final test, then the contribution of the recollection and recollect-to-reject processes for practice testing should be increased on the recollection-

dependent final old/new test (particularly when comparing old and plurality-reversed items), whereas the contribution of familiarity, as was demonstrated in Experiment 1, should be increased on the standard final old/new test and when comparing old and novel lure items on the recollection-dependent final old/new test. Additionally, this should be reflected in the ROCs and zROCs with the ROCs for the comparision of old versus plurality-reversed items in the recollection-dependent test condition being more (or perhaps entirely) linear with an upper x-intercept that is less than 1.0, and the zROCs being U-shaped due to recall dominating the recognition memory judgments and the recollect-to-reject process being used (Rotello et al., 2000). However, if the testing effect results in an increase in the contribution of familiarity in general, then the testing effect may be reduced or attenuated on the recollection-dependent final old/new test, particularly when comparing performance on old versus plurality-reversed items, because familiarity will be less useful for making recognition memory judgments on that format of the final old/new test.

Method

Participants. One hundred (74 females, 26 males with a mean age of 20.8) participants were recruited from the student population of the University of Nevada, Las Vegas via the Department of Psychology's subject pool. For their participation, students were compensated with credit that could be applied to a psychology course. The only restrictions for participation were that one needed to be at least 18 years of age at the time of participation and able to fluently speak and understand English. None of the participants for Experiment 2 participated in Experiment 1.

Materials. The lists (i.e., Lists 1 and 2) from Experiment 1 were also used in

Experiment 2. For the standard final test condition, List 1 was used for the study items during encoding, to create the two-letter stems for the practice test, and for the old items on the final old/new test, and List 2 was used for the new items on the final old/new test. An additional 18 low frequency words (4 to 6 letters in length, 1 to 2 syllables, and concrete nouns) were taken from the MRC Psycholinguistic Database (Kucera & Francis, 1967) for the recollection-dependent final test condition to replace the 18 words from List 1 that could not be turned into its plural form by simply adding an "s" (e.g., ‘glass’; ‘wolf’). For the recollection-dependent final test condition, List 1 (with the 18 replaced words) was used for the study items during encoding where half of the items (i.e., 60 items) were randomly selected to be presented in their plural form and the other half of the items were presented in their singular form (i.e., List 3). List 3 was also used to create the two-letter stems for the practice test. For the final old/new test, half of the singular and plural items from List 3 (i.e., 30 singular and 30 plural words) were randomly chosen to serve as the old items, and the remaining items from List 3 were used to create the plurality-reversed lure items. The plurality-reversed lures were created by changing or reversing the plurality of the chosen studied items (e.g., the studied item of 'marble' was changed to 'marbles' on the final test, whereas the studied item of 'bubbles' was changed to 'bubble' on the final test). In addition, 60 words were randomly chosen from List 2 (half of which were randomly selected to be presented in their plural form) to serve as the novel lure items. Thus, the recollection-dependent final old/new test consisted of 180 items: 60 old items (30 singular and 30 plural old items), 60 plurality-reversed lures (30 singular and 30 plural plurality-reversed lure items), and 60 novel lures (30 singular and 30 plural novel lure items).

Procedure. The basic procedure of Experiment 2 was similar to the one used in the single practice condition of Experiment 1. There were four phases: encoding, distractor, practice, and final old/new test. The major change from Experiment 1 was that the format of the final old/new test was manipulated between-subjects. Participants completed either a standard final old/new test or a recollection-dependent final old/new test. Also, in contrast to Experiment 1, the number of practice sessions was not manipulated in Experiment 2; instead, participants always completed one practice session, regardless of the type of practice. Thus, the type of practice (restudying versus testing) and final test format (standard versus recollection-dependent) were manipulated as between-subjects variables. For the standard final test condition, the procedure was identical to the single practice condition of Experiment 1. For the recollection-dependent final test condition, the procedure was similar to the single practice condition of Experiment 1 except that the studied items during encoding were presented so that half of the items were singular and the other half were plural, and the final test consisted of old items, plurality-reversed lures, and novel lures. Because the singular/plural differences between studied words and similar lures in the recollection-dependent final test condition were somewhat subtle, the instructions for all phases (encoding, practice, and the final test) were tweaked to instruct the participants to pay particular attention to the plurality of the studied words.

Analyses

Similar to Experiment 1, the dependent measures were the proportion of correct responses, *d'* values, ROCs, and parameter estimates, which were calculated using the confidence rating data from the final old/new test. Once again, the practice test data were

not included in any of the analyses for Experiment 2; however, the proportion of correct responses from the practice tests is summarized in the Appendix. The performance measures (proportion of correct responses and *d'* values), ROCs, zROCs and parameter estimates were calculated in the same manner as described and used in Experiment 1. Additionally, because the recollect-to-reject process (R_n) should contribute to performance in the recollection-dependent final test condition, the theoretically based equations of the DPSD model incorporating this process were used to fit the observed confidence rating data for each participant (when comparing old versus plurality-reversed items) to obtain parameter estimates of the familiarity, recollection and recollect-to-reject processes. Using the following theoretically based equations of the DPSD model incorporating a term to represent the recollect-to-reject process (Yonelinas, 1997),

$$(8)\ P(\text{"yes"} \mid \text{studied})_i = R - (\Phi(-d' - c_i))(R - R_n) + (1 - R_n)\Phi(d' - c_i)$$

and

$$(9)\ P(\text{"yes"} \mid \text{new})_i = \Phi(-d' - c_i)(1 - R_n)$$

the DPSD model was fitted to the observed confidence rating data for the comparison of old versus plurality-reversed items. The process for fitting these equations to the observed data was the same as the procedure described and used in Experiment 1 (i.e., a log-likelihood estimation method using the solver function in Microsoft Excel). Thus, for Experiment 2, for each participant's data, the equations of the DPSD model (using the appropriate equations) were fitted to the observed confidence rating data to obtain parameter estimates of the recollection and familiarity processes when comparing old versus novel lure items and the parameter estimates of the recollection, recollect-to-reject and familiarity processes when comparing old versus plurality-reversed lure items. The

group average of the parameter estimates is what was analyzed using between-subjects ANOVAs to determine how the processes involved in making the recognition memory judgments on the final test were affected by the experimental manipulations (i.e., for Experiment 2, restudying versus practice testing and standard versus recollection-dependent final test formats).

Results and Discussion

In the same manner as Experiment 1, proportion of correct responses, *d'* values, ROCs and parameter estimates were calculated using the confidence ratings from the final old/new test. Also, as in Experiment 1, an alpha level of .05 was used to determine statistical significance for all analyses. In contrast to Experiment 1, for the data from the recollection-dependent final test condition, *d*' values, ROCs and parameter estimates were calculated for both the comparison between old versus novel lure items and for the comparison between old versus plurality-reversed lure items. Thus, overall proportion of correct responses, *d*' values for old versus novel lure items (old-novel *d'*), and parameter estimates of familiarity and recollection for old versus novel lure items (old-novel familiarity and old-novel recollection) were each analyzed using a 2 (practice condition: restudying versus testing) x 2 (final test format: standard versus recollection-dependent) between-subjects ANOVA. Additionally, for the comparison of old versus plurality-reversed lure items in the recollection-dependent final test condition, *d'* values (old-similar *d'*) and the parameter estimates of the familiarity (old-similar familiarity), recollection (old-similar recollection) and recollect-to-reject processes were each analyzed using a 2-way (practice condition: restudying versus testing) between-subjects ANOVA.

The results for performance in terms of both the proportion of correct responses data and old-novel *d'* values demonstrated the testing effect, replicating the findings of Experiment 1 (see Figures 7 and 8). That is, practice testing significantly enhanced performance (i.e., higher proportion of correct responses and old-novel *d'* values) on the final old/ new test compared to restudying, regardless of the format of the final test, $F(1, 99) = 18.711$, $p < .001$, $\eta_p^2 = .163$ and $F(1, 99) = 17.975$, $p < .001$, $\eta_p^2 = .158$, respectively for proportion of correct responses and old-novel *d'*. Additionally, a significant main effect of final test format was found demonstrating enhanced long-term retention (i.e., significantly higher proportion of correct responses and old-novel *d'* values) for the standard final test format compared to the recollection-dependent final test format, regardless of the type of practice, $F(1, 99) = 63.973$, $p < .001$, $\eta_p^2 = .400$ and $F(1, 99) = 18.100$, $p < .001$, $\eta_p^2 = .159$, respectively for proportion of correct responses and old-novel *d'*. Finally, no interaction between practice condition and final test format was found for either proportion of correct responses or old-novel *d*' values, $F(1, 99) = 0.010$, $p = .922$, $\eta_p^2 = .000$ and $F(1, 99) = 0.182$, $p = .671$, $\eta_p^2 = .002$, respectively. Planned comparisons for both accuracy measures were conducted using independent-samples *t*-tests. Enhanced performance with both proportion of correct responses and old-novel *d'* values was found for practice testing compared to restudying with both the standard final test format, $t(48) = 2.665$, $p = .010$ and $t(48) = -2.469$, $p = .017$, and recollection-dependent final test format, $t(48) = 3.632$, $p = .001$ and $t(48) = 3.673$, $p = .001$. Additionally, significantly higher proportion of correct responses and old-novel *d'* values was found for the standard final test format compared to the recollection-dependent final test format for both practice testing, $t(48) = 5.613$, $p < .001$ and $t(48) = 2.586$, $p = .013$,

and restudying, $t(48) = 5.698$, $p < .001$ and $t(48) = 3.480$, $p = .001$. These findings replicate those found in Experiment 1 demonstrating that practice testing enhances later memory performance compared to restudying, and extend the findings of Experiment 1 by showing that practice testing can benefit later memory performance on both standard and recollection-dependent final test formats.

The parameter estimates of familiarity and recollection from modeling the confidence rating data for old versus novel lure items using the DPSD model produced similar results to those described in Experiment 1 for one practice session. That is, the patterns of accuracy performance just described led to differences in the old-novel parameter estimates with these differences being reflected in the ROCs and zROCs in a manner consistent with the DPSD model (see Figures 9-12). Specifically, the analysis of the old-novel familiarity parameter estimates (see Figure 9) showed a significant increase in the contribution of old-novel familiarity for practice testing compared to restudying, regardless of the final test format, $F(1, 99) = 9.483$, $p = .003$, $\eta_p^2 = .090$. A significant increase in the contribution of old-novel familiarity was found with the standard final test format compared to the recollection-dependent final test format, regardless of the type of practice, $F(1, 99) = 11.788$, $p = .001$, $\eta_p^2 = .109$. No interaction between practice condition and final test format was found, $F(1, 99) = 0.002$, $p = .966$, $\eta_p^2 = .000$. Planned comparisons conducted using independent-samples *t*-tests showed that practice testing (compared to restudying) significantly increased the contribution of old-novel familiarity for the recollection-dependent final test format, $t(48) = 2.450$, $p = .018$, while for the standard final test format, practice testing (compared to restudying) led to a marginally significant increase in the contribution of old-novel familiarity, $t(48) = 1.970$, $p = .055$.

Additionally, the standard final test format (compared to the recollection-dependent final test format) significantly increased the contribution of old-novel familiarity for both practice testing, $t(48) = 2.225$, $p = .031$, and restudying, $t(48) = 2.684$, $p = .010$.

The analysis of the old-novel recollection parameter estimates (see Figure 10) did not show a significant difference between the practice conditions, $F(1, 99) = 2.069$, $p = .154$, $\eta_p^2 = .021$, nor the final test formats, $F(1, 99) = 0.477$, $p = .505$, $\eta_p^2 = .005$. There also was not an interaction between practice condition and final test format, $F(1, 99) = 0.022$, $p = .883$, $\eta_p^2 = .000$. The planned comparisons conducted using independent-samples *t*-tests showed no difference in the contribution of old-novel recollection between the final test formats for neither the testing group, $t(48) = 0.543$, $p = .590$, nor the restudying group, $t(48) = 0.395$, $p = .695$. Also, no significant difference in the contribution of old-novel recollection was found between practice testing and restudying for neither the standard final test format, $t(48) = 1.165$, $p = .250$, nor the recollection-dependent final test format, $t(48) = 0.881$, $p = .383$. Taken together, the results from the old-novel parameter estimates replicate the findings from the single practice condition in Experiment 1, once again suggesting that enhanced performance with practice testing is generally due to an increase in the contribution of familiarity.

Generally, the increase in familiarity (but not in recollection) for old versus novel lure items just described was reflected in the ROCs and zROCs for each condition in a manner that is consistent with what the DPSD model would predict (see Figures 11 and 12). As the DPSD model would predict for tests of item recognition when comparing old and novel lure items (Parks & Yonelinas, 2008), overall the ROCs were curvilinear and asymmetrical along the negative diagonal reflecting the contribution of both familiarity

and recollection, and the zROCs were approximately linear. Additionally, the ROCS for practice testing and the standard final test format conditions were somewhat more curvilinear and symmetrical (along the negative diagonal) with a more gradual increase towards 1, 1 for the practice testing conditions (as compared to the restudying conditions) and the standard final test format conditions (as compared to the recollection-dependent final test format conditions) reflecting the increase in old-novel familiarity (but not old-novel recollection) for those conditions. Taken together, the results for the *d′* values, parameter estimates and ROCs for the comparison of old versus novel lure items are consistent with those seen in Experiment 1 with a single practice condition, again supporting the notion that the DPSD model can be a useful method for investigating the testing effect.

However, the results from the old versus plurality-reversed lure items for the recollection-dependent final test condition suggest that practice testing may not always be beneficial to performance (see Figure 13-18). When examining performance using the old-similar *d'* values (see Figure 13), no significant difference was found between practice testing and restudying, $F(1, 49) = 1.712$, $p = .197$, $\eta_p^2 = .034$. The parameter estimates of the familiarity, recollection and recollect-to-reject processes from modeling the confidence rating data for old versus plurality-reversed lure items using the DPSD model were consistent with the old-similar *d′* results (see Figures 14-16). That is, no significant differences between the practice conditions were found for neither the parameter estimates of old-similar familiarity, $F(1, 49) = 1.779$, $p = .189$, $\eta_p^2 = .036$, old-similar recollection, $F(1, 49) = 0.159$, $p = .692$, $\eta_p^2 = .003$, nor recollect-to-reject, $F(1, 49) = 0.581$, $p = .450$, $\eta_p^2 = .012$. While there was no difference between the practice

conditions in old-similar *d′* nor any of the old-similar parameter estimates, the old-similar ROCs and zROCs did reflect performance and the parameter estimates in a manner that is consistent with the DPSD model for tests of item recognition that contain similar lures (see Figures 17 and 18). That is, overall the old-similar ROCs were linear and had an upper x-intercept less than 1.0 with the zROCS being slightly U-shaped, reflecting the use of the recollect-to-reject process. Taken together, the performance data and parameter estimates for old versus plurality-reversed lure items on the recollection-dependent final test do not demonstrate any differences between the practice conditions, which may suggest that practice testing does not aid memory (in comparison to restudying) when performance is dependent on making a discrimination between items on a relatively small sets of features (i.e., the plurality of the word). However, performance was very low in both practice conditions. Thus, the lack of differences between the practice conditions may be due to low performance (due to the difficulty of the recollection-dependent final test format), and not necessarily that practice testing does not benefit performance when discriminating between items based on a relatively small sets of features. Furthermore, even though there were no differences between the practice conditions for old-similar comparisons, the results were still consistent with the DPSD model and thus provide further support for the conclusion that the DPSD model can be used to investigate the testing effect.

CHAPTER 6

GENERAL DISCUSSION

The goal of this dissertation was to test whether the testing effect could be examined using the DPSD model (Yonelinas, 1994). Two experiments were conducted where practice testing was compared to restudying on tests of long-term item recognition memory to assess the utility of the DPSD model for examining the testing effect. Generally, both experiments revealed performance benefits of practice testing over restudying on both final test formats, with a greater benefit of practice testing found in Experiment 1 with three practice sessions compared to a single practice session. This replicates previous findings in the testing effect literature demonstrating that multiple practice tests leads to greater performance than a single practice test, and practice testing, generally, leads to greater long-term memory than restudying (for reviews, see Roediger & Butler, 2011; Rodegier & Karpicke, 2006b). Importantly, these findings also demonstrate that the benefits of practice testing were reflected in the processes involved in making the recognition memory judgments (when there was a benefit of practice testing to performance) in a manner consistent with the DPSD model. Specifically, the enhancement of memory with practice testing also led to increases in the contribution of familiarity when compared to restudying (on both final test formats), with greater increases seen with three practice tests. Furthermore, with three practice tests, there was also an increase in the contribution of recollection. These findings demonstrate that the DPSD model can be used to investigate the benefits of retrieval practice (e.g., through practice testing) on long-term recognition memory under a variety of conditions (i.e., manipulations of the number of practice sessions and format of the final recognition

memory test).

Overall, based on the findings of this study, the DPSD model may provide a useful approach for examining the testing effect, both in terms of the experimental conditions that influence the testing effect as well as theoretical explanations of the testing effect. The majority of the testing effect research has focused on investigating the factors that influence the testing effect and applications of the testing effect, and because research focusing on theoretically understanding of the effect has been limited, the mechanisms responsible for the effect are not well understood. Furthermore, the testing effect is not always found when a recognition final test is used (Chan & McDermott, 2007; Roediger & Karpicke, 2006). Thus, based on the findings of this study, the DPSD model may be a useful approach for future testing effect research. The DPSD model could help to provide insight into why the testing effect is not consistently shown when a recognition final test is used by allowing for the examination of the processes underlying recognitiong memory judgments and performance. Understanding the processes underlying recognition memory, using the DPSD model, could also be valuable to examining and adding to the theoretical explanations for the effect, which in turn could lead to more informed recommendations about the application of the tesing effect in terms of both when and how retrieval practice should be used to enhance learning and long-term retention.

This study extends the previous studies conducted by Chan and McDermott (2007) and Verkoeijen et al. (2011) by examining the influence of retrieval practice (using practice testing) on the processes of recollection and familiarity when practice testing (in comparison to restudying) actually led to benefits in memory performance.

Additionally, this study examined the benefits of practice testing to memory performance with multiple practice sessions, a mainpulation known to increase the magnitude of the testing effect. In contrast to these previous studies, which both found that practice testing increased recollection (but did not influence familiarity), the findings of this dissertation demonstrated that practice testing increases the contribution of familiarity and only increased the contribution of recollection (in addition to familiarity) with multiple practice sessions. Based on the dual-process literature, there are a few possibilities for these different findings. First, the increases in recollection that were found by Chan and McDermott (2007) and Verkoeijen et al. (2011) were on an immediate test of recognition memory (when there was no benefit of practice testing). Previous dual-process research has shown that across short-term retention intervals, the forgetting rate for familiarity-based judgments is greater and more rapid than the forgetting rate for recollection-based judgments (i.e., familiarity decreases rapidly while recollection remains relatively unaffected), whereas for long-term retention intervals (i.e., as the delay between study and test increases) as was used in this study, the forgetting rates for both familiarity- and recollection-based judgments are similar (i.e., both familiarity and recollection decrease significantly, at comparable rates) (see Yonelinas, 2002 for a review). Thus, it is possible that the reason that Chan and McDermott (2007) and Verkoeijen et al. (2011) demonstrated increases in recollection, while this study generally demonstrated increases in familiarity could be due to the difference in the retention intervals between encoding/practice and the final test. Second, in the Chan and McDermott (2007) and Verkoeijen et al. (2011) studies, recognition memory was assessed using source and exclusion tests, whereas in this study recognition memory was assessed using tests of

item recognition. This distinction in how recognition memory was assessed is important because previous dual-process research has shown that performance on tests of item recognition relies on a combination of recollection and familiarity, whereas performance on source and exclusion tests relies primarily (although not exclusively) on recollection (e.g., Yonelinas, 1997; 1999). Therefore, it is possible that Chan and McDermott (2007) and Verkoeijen et al. (2011) found that practice testing influences recollection but not familiarity due to the way that recognition memory was assessed. Finally, a third possibility for the differences between the previous studies and this study is the difference in the measurement methods used to estimate recollection and familiarity. Chan and McDermott (2007) estimated recollection and familiarity using the process dissociation and remember/know procedures, Verkoeijen et al. (2011) used the process dissociation procedure, and finally in this study the ROCs procedure based on the equations of the DPSD model was used. While generally it has been demonstrated that these various measurement methods lead to converging results in terms of the process estimates based on a variety of experimental manipulations (Yonelinas, 2001b; 2002), it is possible that the use of different measurement methods led to the differing results. However, regardless of whether this is truly a plausible explanation for the differences in the results between this study and previous studies, it would be beneficial for future research to examine the testing effect (i.e., benefits of retrieval practice on tests of long-term memory) using other measurement methods (i.e., the process dissociation and remember/know procedures) on various tests of recognition memory (e.g., item recognition and source memory).

While the results of this study support the notion that the DPSD model can be

used to investigate the testing effect and manipulations that influence the magnitude of the testing effect (i.e., number of practice sessions), it is important to discuss the findings from the recollection-dependent final test format, where a benefit of practice testing (in comparison to restudying) was not found. As a reminder, when comparing old items versus plurality-reversed lures, no significant differences between practice testing and restudying were found for old-similar *d'* values nor any of the parameter estimates; however, despite the lack of performance and parameter estimate differences between the practice conditions, the shapes of the old-similar ROCs and zROCs were in line with the patterns that the DPSD model would predict. The lack of differences between the practice conditions with the comparison between old and plurality-reversed items along with the increase in familiarity for practice testing (compared to restudying) seen in Experiment 1 and with the comparison between old and novel lure items in Experiment 2 could be interpreted as support for the notion that practice testing results in an increase in the contribution of familiarity in general. This notion would explain why there was no benefit of practice testing on the recollection-dependent final test format when comparing old and plurality-reversed lure items as the ability to discriminate between old and plurality-reversed lure items would rely heavily, if not exclusively, on recollection processes as both type of items would be familiar and thus make the familiarity process less useful in making these type of recognition memory judgments. Furthermore, if practice testing increases familiarity in general, then it may suggest that practice testing would not be a superior practice strategy to restudying when memory performance is based on the ability to discriminate between items based on a relatively small sets of features (e.g., in this study, the ability to discriminate between old and similar lure items based on the plurality

of the word).

However, there are a couple of patterns in the results of this study that suggest that it may be premature to conclude that practice testing (with long-term retention) increases the contribution of familiarity in general, and thus practice testing would not be beneficial to long-term recognition memory on tests that heavily rely on recollection. First, performance for both practice conditions on the recollection-dependent final test format was pretty low, particularly when looking at the comparison between old versus plurality-reversed items. Overall, performance on the recollection-dependent final test was 62% and 56% for practice testing and restudying, respectively. Furthermore, *d'* for old versus plurality-reversed lure items was 0.47 and 0.36 for practice testing and restudying, respectively. Finally, when looking at the proportion of correct responses from the recollection-dependent final test for plurality-reversed lures only, performance was approximately 45% for both practice conditions. Thus, as evident in the numbers just presented, performance was near chance for both practice conditions, probably due to the difficulty of the format of the recollection-dependent test, particularly the difficulty in discriminating between old and plurality-reversed lure items based on a small set of features (i.e., the presence of an "s" or not during encoding compared to what was presented on the final test). Therefore, it is possible that deep encoding of the stimulus features is needed to discriminate between items on a small set of features (e.g., between old and plurality-reversed lures), and if the stimuli are not adequately encoded during initial study then practice testing may not benefit performance (at least with a single practice test). Second, an increase in recollection (in addition to familiarity) with practice testing (in comparison to restudying) was found in Experiment 1 when three practice

sessions were used. It is possible that the increase in recollection seen with the three practice tests condition could be due to an increase in study duration (i.e., repeating items using distributed practice) and generation (i.e., generating a word at the time of study compared to reading the word) as both encoding manipulations tend to lead to slightly larger increases in recollection than familiarity (Yonelinas, 2002). However, the increase in study duration also occurred for the three session restudying condition and generation occurred with all practice testing conditions. Furthermore, a recent study conducted by Karpicke and Zaromb (2010) compared generation and retrieval practice, and found that retrieval practice enhanced future retention to a greater extent than generation, demonstrating that the testing effect and generation effect are distinct effects. Taken together, the low performance on the recollection-dependent final test format and increase to both familiarity and recollection with three practice tests suggest that the conclusion that practice testing increases familiarity in general may be incorrect (or at the very least premature), and instead practice testing may influence both familiarity and recollection but this was not evident in Experiment 2 when comparing old versus plurality-reversed items because of low performance on the recollection-dependent final test. Thus, to truly answer the question of whether practice testing increases familiarity in general or increases recollection and/or familiarity based on the format of the final test, further research is needed. One way to further examine this question would be to implement three practice sessions with a recollection-dependent final test as this would enhance performance overall and possibly lead to an increase in recollection for practice testing (but not restudying) as was demonstrated in Experiment 1. Another method for further assessing this question would be to compare practice testing and restudying on

other tests that are recollection-dependent, such as tests of associative and source memory. This would allow for the examination of whether practice testing simply does not lead to an increase in recollection and thus is not beneficial to recollection-dependent tests, or instead (and perhaps more likely) that practice testing just is not beneficial on recollection-dependent tests that rely on such a small, fine discrimination as is the case with a plurality-reversed task.

Finally, the results of the current study may also contribute to possible explanations of the testing effect. While the current theoretical explanations of the testing effect discussed earlier were not directly tested in this dissertation, the findings could be interpreted as support for the cue diagnosticity perspective (Karpicke & Blunt, 2011; Karpicke & Smith, 2012; Karpicke & Zaromb, 2010). As a reminder, the basic notion behind this perspective is that retrieval practice benefits memory by enhancing the diagnostic value of retrieval cues as opposed to an increase or addition to encoded features (Karpicke & Smith, 2012; Karpicke & Zaromb, 2010). The findings of the current study could be seen as support for the cue diagnosticity explanation of the testing effect in that practice testing only enhanced performance to a greater extent than restudying when practice testing increased the diagnostic value of the retrieval cues leading to the ability to discriminate between the target and other potential candidates (i.e., when discriminating between old and novel items but not when discriminating between old and plurality-reversed items). It seems difficult, though not impossible, to explain the findings of this study in terms of other explanations of the testing effect based on elaboration as presumably elaboration would have occurred during practice testing for both final test format conditions, and thus a benefit of practice testing should have been

seen even when comparing old versus plurality-reversed items in Experiment 2. Further research is needed to identify the causal mechanism(s) underlying the testing effect as well as to tease apart the current theoretical explanations that have been proposed for the testing effect. Nevertheless, this dissertation does demonstrate that the DPSD model can be used to investigate the testing effect, and thus may provide a useful method for assessing the various theoretical explanations of the testing effect.

CHAPTER 7

LIMITATIONS AND FUTURE DIRECTIONS

One of the limitations of this dissertation may lie in the inability to directly compare the finding from this study to those conducted by Chan and McDermott (2007) and Verkoeijen et al. (2011) due to procedural difference between the studies. Those previous studies included short retention intervals that typically do not lead to a testing effect, whereas the current study included a longer retention interval because the majority of studies have shown that the testing effect is typically found with long retention intervals. In addition, in this study, recollection and familiarity were estimated using a direct measurement method (i.e., the ROCS procedure based on the equations of the DPSD model). These changes make it difficult to determine whether the differences in the findings between the studies occurred because there are differences in how practice testing influences recollection and familiarity based on whether enhancement to memory is observed, or due to these procedural difference between the studies (and how these differences influence the contribution of recollection and familiarity to memory judgments). The use of other measurement methods (i.e., the process-dissocaition and remember/know procedures) could be employed in a future study with both short- and long-term retention intervals to directly compare these various measurement methods. However, the ROCs procedure is one of the most direct methods for estimating recollection and familiarity, and previous dual-process research has shown that the various measurement methods typically led to similar results across a variety of experimental manipulations (Yonelinas, 2001b; 2002). Based on this, it seems likely that the difference in the findings between the previous studies and this one are due to the

retention interval used. Nevertheless, it would be interesting to determine whether the different measurement methods would produce similar results with a testing effect paradigm.

Another important finding to note is the consistent observation in this study that practice testing primarily influenced the contribution of familiarity; the three practice tests condition in Experiment 1 was the only condition where the parameter estimates showed an increase in both familiarity and recollection. It is unclear whether the increased contribution of recollection observed with three practice tests had to do with the combination of increased studying duration and use of generation in that condition (both of which are factors that have been shown to increase recollection slightly more than familiarity in the dual-process literature; Yonelinas, 2002) or because the benefits of practice testing can lead to increases of recollection but were not observed when comparing old and plurality-reversed items in Experiment 2 because of other factors (e.g., chance performance, use of a single practice session, etc.). Future research could further examine this issue by attempting to replicate the increase in recollection with three practice tests as well as by examining whether multiple practice sessions (e.g., three practice sessions) with a recollection-dependent final test format would lead to a different pattern of a results than those found in Experiment 2. Additionally, this issue could be addressed by using other types of recognition memory tests that primarily rely on recollection, such as source and exclusion tests like those used in the Chan and McDermott (2007) and Verkoeijen et al. (2011) studies.

CHAPTER 8

CONCLUSION

In summary, this dissertation demonstrated that the DPSD model offers a useful approach for investigating the testing effect. The findings from both experiments demonstrated that the benefits of retrieval practice (through practice testing) can be explained by increases of the processes involved in making recognition memory judgments in a manner that is in line with the DPSD model. Importantly, the current study went beyond simply demonstrating that the DPSD model can be used to investigate the testing effect by examining both the basic testing effect as well as an important factor (i.e., the number of practice sessions) that influences the magnitude of the testing effect. Based on the findings from this dissertation, the DPSD model could be used to further understand the testing effect in terms of the processes responsible for the effect and the experimental manipulations that increase or reduce the benefits of practice testing as well as the testing situations that lead to the testing effect. Furthermore, the DPSD model may also provide a useful and informative method for examining the mechanisms underlying the testing effect and the various theoretical explanations that have been proposed to account for the effect.

REFERENCES

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8*, 463-470.

Bacon, F. (2000). *Novum organum* (L. Jardine & M. Silverthorne, Eds.). Cambridge, England: Cambridge University Press. (Original work published in 1620).

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. S. Nairne (Ed.), *The foundation of remembering: Essays in honor of Henry L. Roediger, III* (pp. 83-105). New York: Psychology Press.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124(2)*, 137-160.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(5)*, 1118-1133.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273-281.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitve error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(4)*, 918-928.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19(4/5)*, 514-527.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36(3)*, 604-616.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(6)*, 1563-1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(6)*, 1547-1552.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34(2)*, 268-276.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20(6)*, 633-642.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological*

Bulletin, 132(3), 354-380.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(2)*, 431-437.

Darley, C. F., & Murdock, B. B., Jr. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91(1)*, 66-73.

Estes, W. K. (1960). Learning theory and the new "mental chemistry." *Psychological Review, 67,* 207-223.

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *The Quarterly Journal of Experimental Psychology, 60*, 991-1004.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6(40).*

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13(1)*, 8-20.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81(3)*, 392-399.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10*, 562-567.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*, 340-344.

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

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relationship between conscious and unconscious (automatic) influences: A declaration of independence. In J. Cohen & J. W. Schooler (Eds.), *Scientific Approaches to the Questions of Consciousness* (pp. 13-47). Hillsdale, NJ: Earlbaum.

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

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19(4/5)*, 528-558.

Kapucu, A., Macmillan, N. A., & Rotello, C. M. (2010). Positive and negative remember judgments and ROCs in the plurals paradigm: Evidence for alternative decision strategies. *Memory & Cognition, 38(5)*, 541-554.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772-775.

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151-162.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*, 17-29.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*, 227-239.

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 701-722.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American*

English. Providence: Brown University Press.

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2008). Test-enhanced learning in medical education. *Medical Education, 42*, 959-966.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29(3)*, 210-212.

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition, 15*, 257-280.

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

Mandler, G., & Rabinowitz, J. C. (1981). Appearance and reality: Does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory, 7(2)*, 79-90.

McDaniel, M. A. (2007). Transfer: Rediscovering a central concept. In H. L. Roediger, III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *The science of memory: Concepts*. New York: Oxford University Press.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007), Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19(4/5)*, 494-513.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*, 192-201.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, 11, 371-385.

McDaniel, M. A., McDermott, K. B., Agarwal, P. K., & Roediger, H. L., III (2008, June). Test-enhanced learning in the classroom: The Columbia Middle School project, year 2. Poster presented at the meeting of the Institute of Education Sciences Research, Washington, D.C.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14(2)*, 200-206.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*, 99-113.

Moscovitch, M., & Craik, F. I. M. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 447-458.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory, 10(5/6)*, 389-395.

Parks, C. M., Murray, L. J., Elfman, K., & Yonelinas, A. P. (2011). Variations in recollection: The effects of complexity on source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(4)*, 861-873.

Parks, C. M., & Yonelinas, A. P. (2008). Theories of recognition memory. In H. L. Roediger, III (Ed.), *Cognitive Psychology of Memory* (Vol. 2 of Learning and memory: A comprehensive reference, pp. 389-416). Oxford: Elsevier.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback

facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(1)*, 3-8.

Pashler, H., Zarrow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1051-1057.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335.

Pyc, M.A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 28(3)*, 737-746.

Roediger, H. L., III, Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies and D. B. Wright (Eds.), *New Frontiers in Applied memory* (pp.13-49) Brighton, U.K.: Psychology Press.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20-27.

Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17(3)*, 249-255.

Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1(3)*, 181-210.

Roediger, H. L., III, McDaniel, M. A., McDermott, K. B., & Agarwal, P. K. (2007, November). Test-enhanced learning in the classroom: The Columbia Middle

School project. Poster presented at the meeting of the Psychonomic Society, Long Beach, CA.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(1)*, 233-239.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43*, 67-88.

Sun, J., & McDaniel, M. A. (2008, November). The testing effect: Experimental evidence from a college course. Poster presented at the meeting of the Midstates Consortium for Math and Science, Chicago, IL.

Szpunar, K. K., McDermott, K. B., Roediger, H. L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6)*, 1392-1399.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4(3)*, 210-221.

Tulving, E. (1967). The effects of presentation and recall of material in free recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175-184.

Tulving, E. (1974). Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American Scientist, 62(1)*, 74-82.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist, 26*, 1-12.

Verkoeijen, P. P. J. L., Tabbers, H. K., & Verhage, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology, 58(6)*, 490-498.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571-580.

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3(4)*, 240-245.

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

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

Yonelinas, A. P. (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. (1999). The contribution of recollection and familiarity to recognition and source memory judgments: A formal dual-process model and an ROC analysis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 1415-1434.

Yonelinas, A. P. (2001a). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society, 356*,

1363-1374.

Yonelinas, A. P. (2001b) Consciousness, control, and confidence: The 3 Cs of Recognition Memory. *Journal of Experimental Psychology: General, 130(3)*, 361-379.

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

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

Yonelinas, A. P., & Levy, B. J. (2002). Dissociating familiarity from recollection in human recognition memory: Differences of forgetting over short retention intervals. *Psychonomic Bulletin & Review, 9(3)*, 575-582.

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

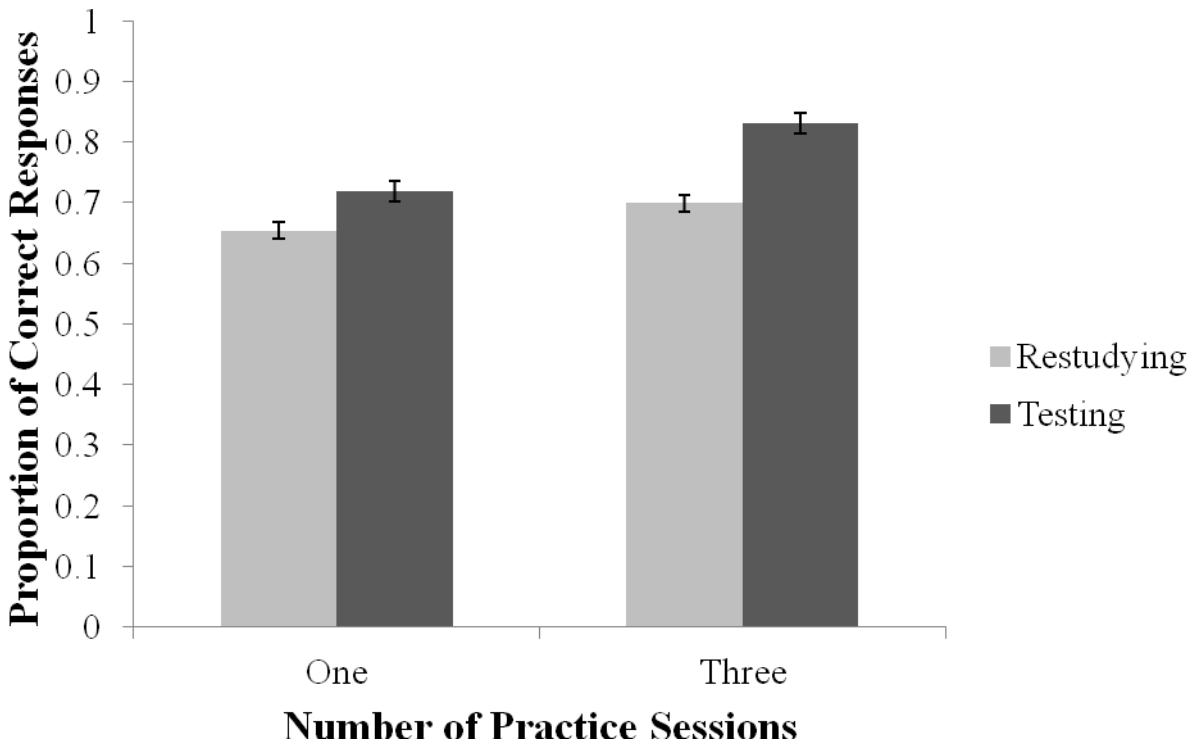

Figure 1. Proportion of correct responses for Experiment 1.

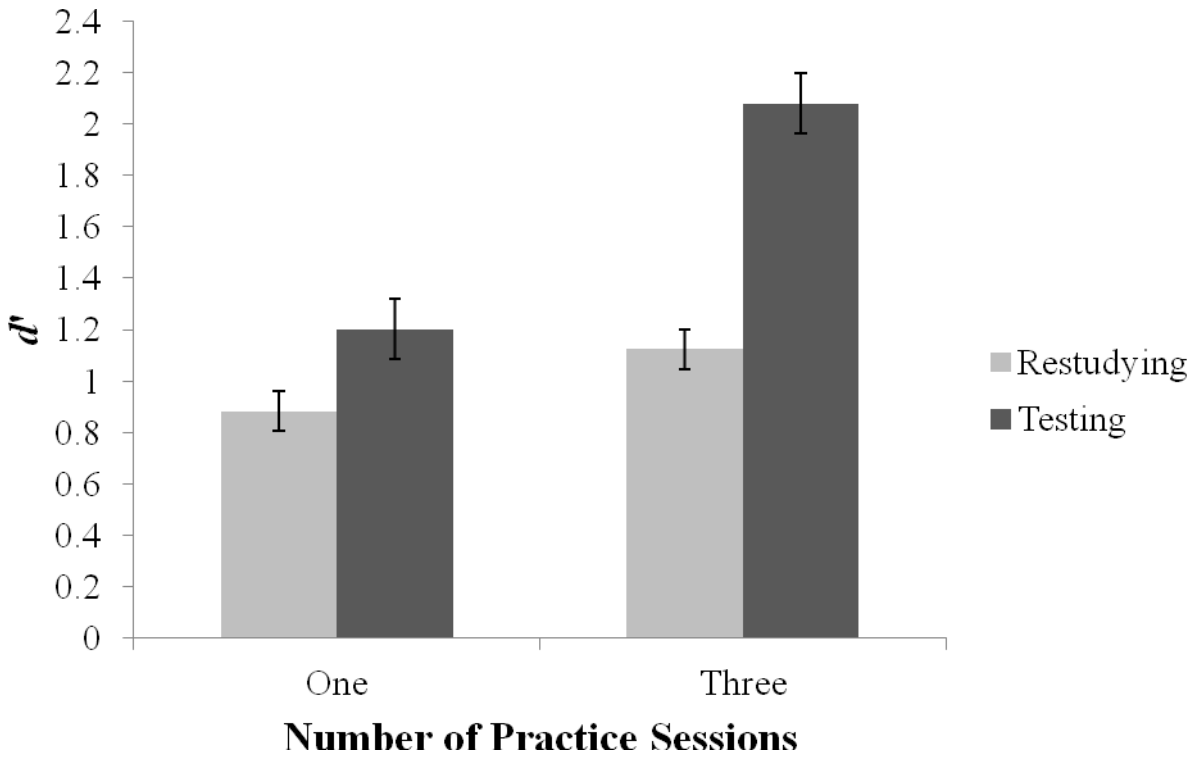

Figure 2. d' values for Experiment 1.

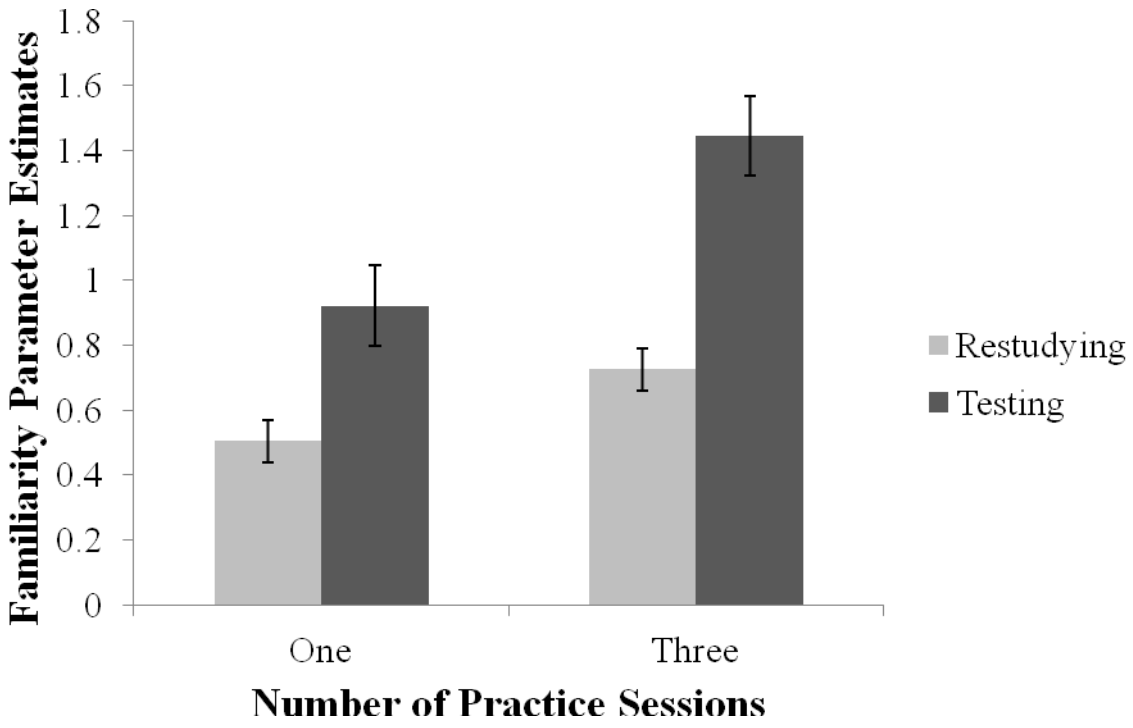

Figure 3. Familiarity parameter estimates for Experiment 1.

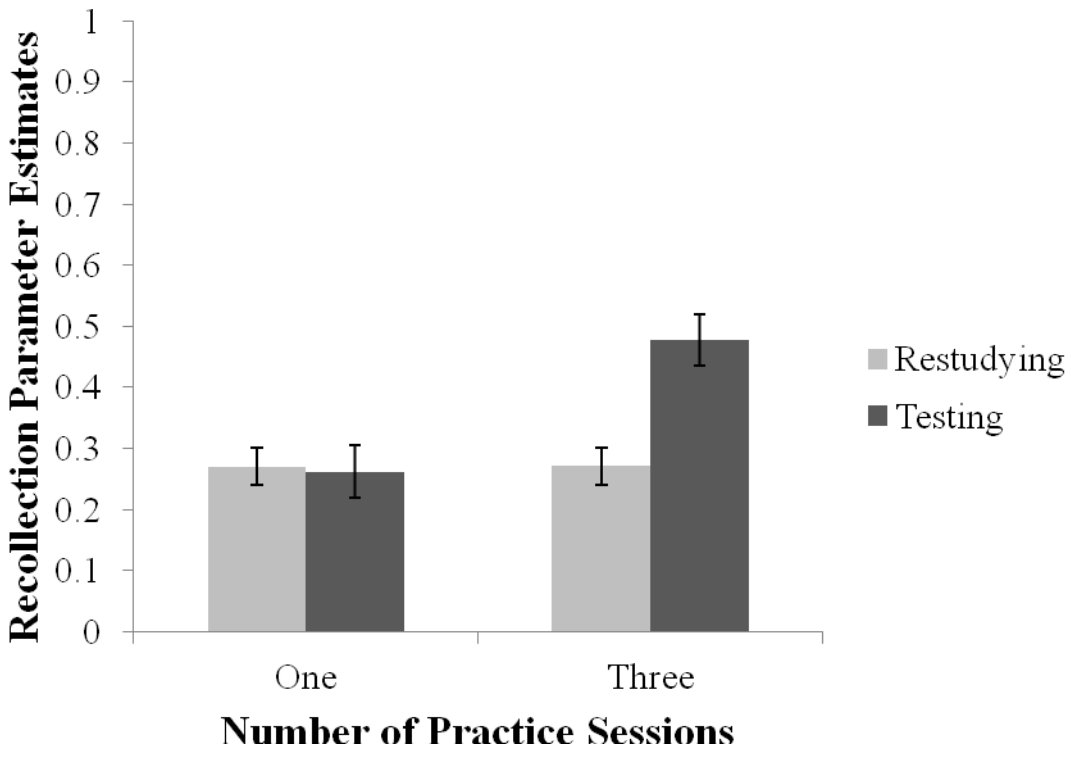

Figure 4. Recollection parameter estimates for Experiment 1.

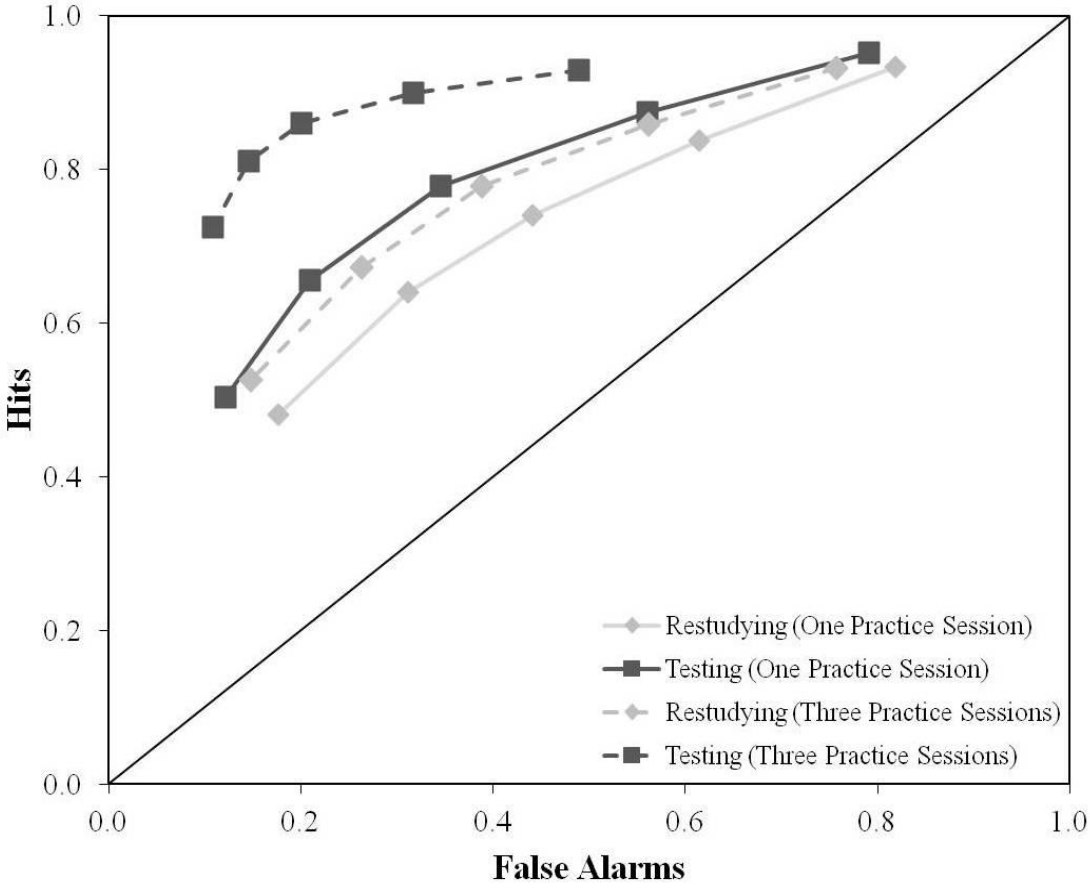

Figure 5. ROCs for Experiment 1.

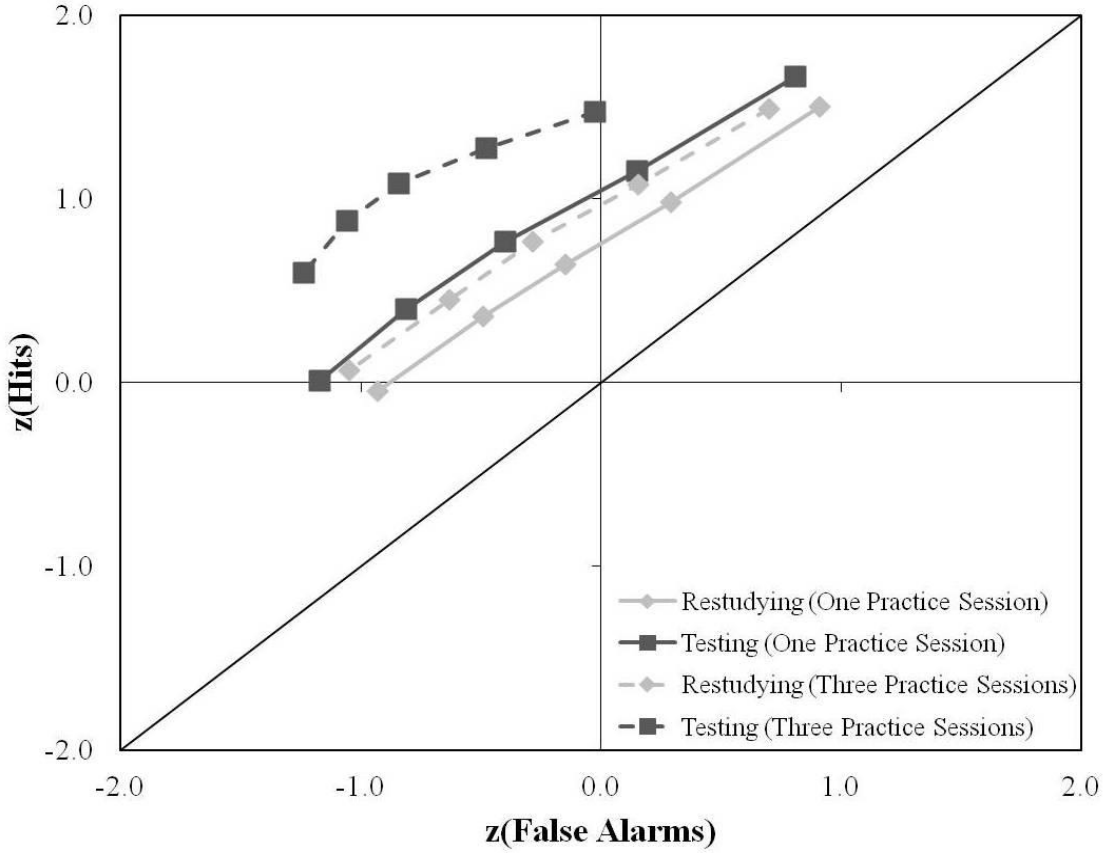

Figure 6. zROCs for Experiment 1.

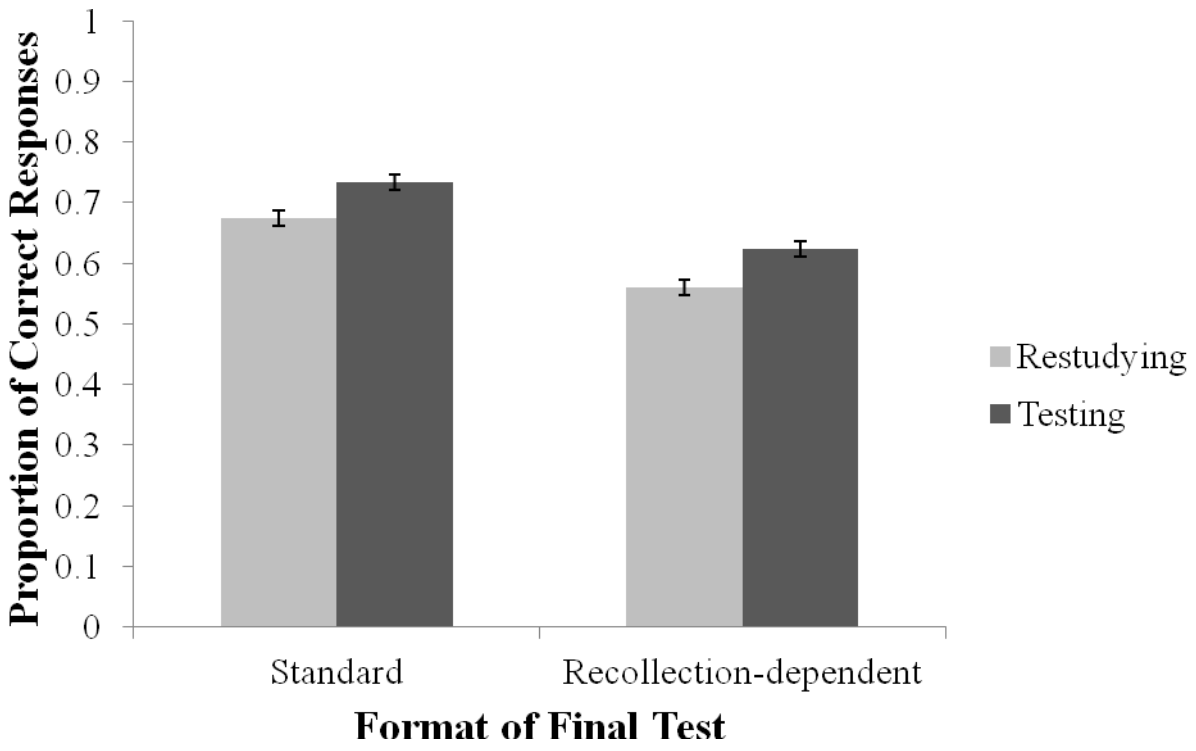

Figure 7. Proportion of correct responses for Experiment 2.

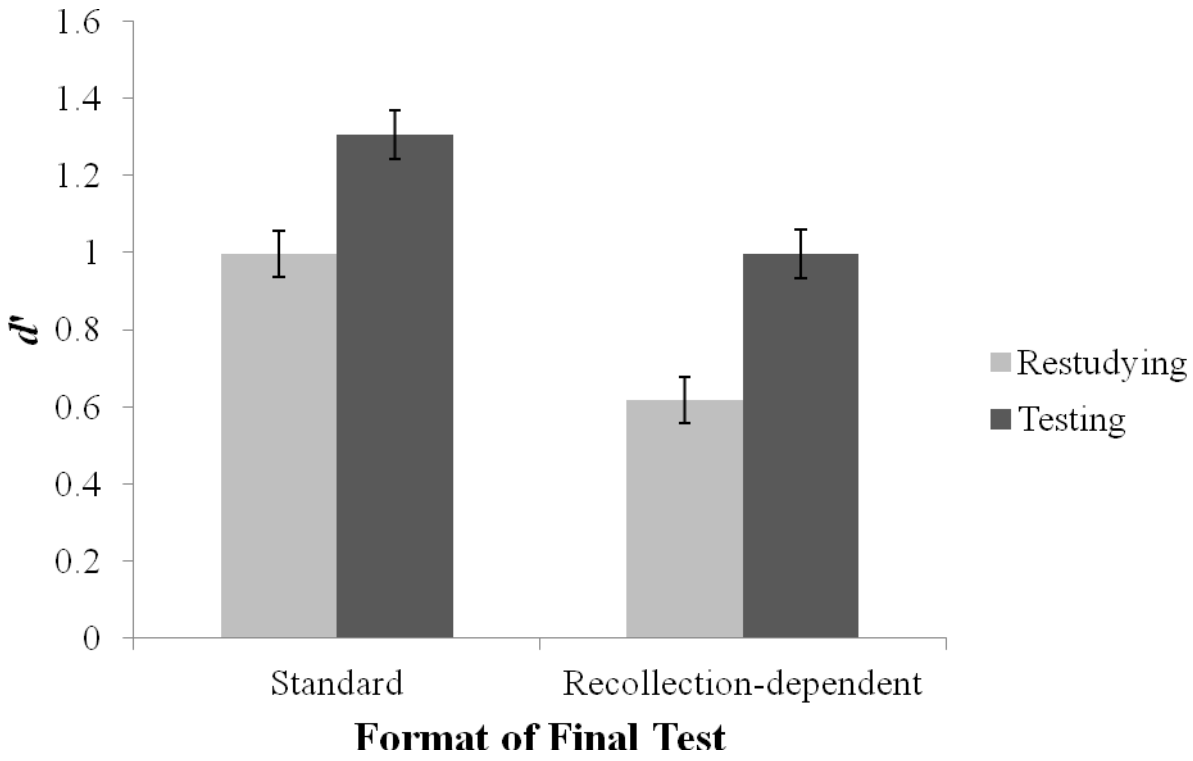

Figure 8. d' values for old versus novel items for Experiment 2.

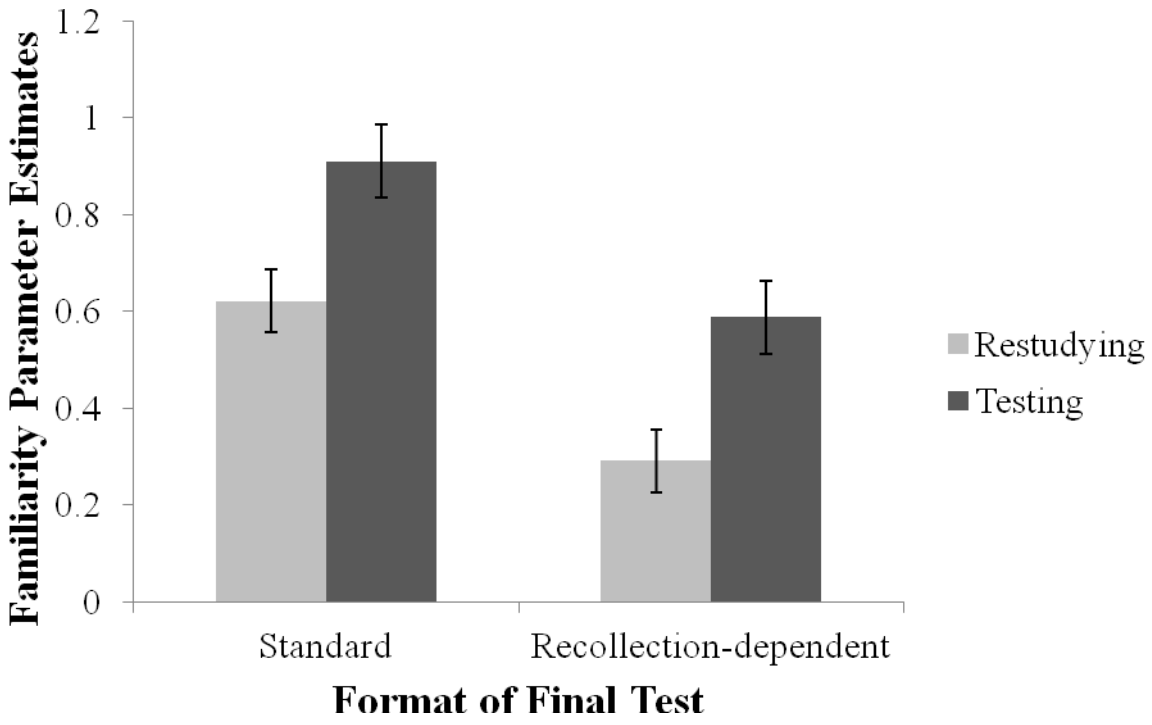

Figure 9. Familiarity parameter estimates for old versus novel items for Experiment 2.

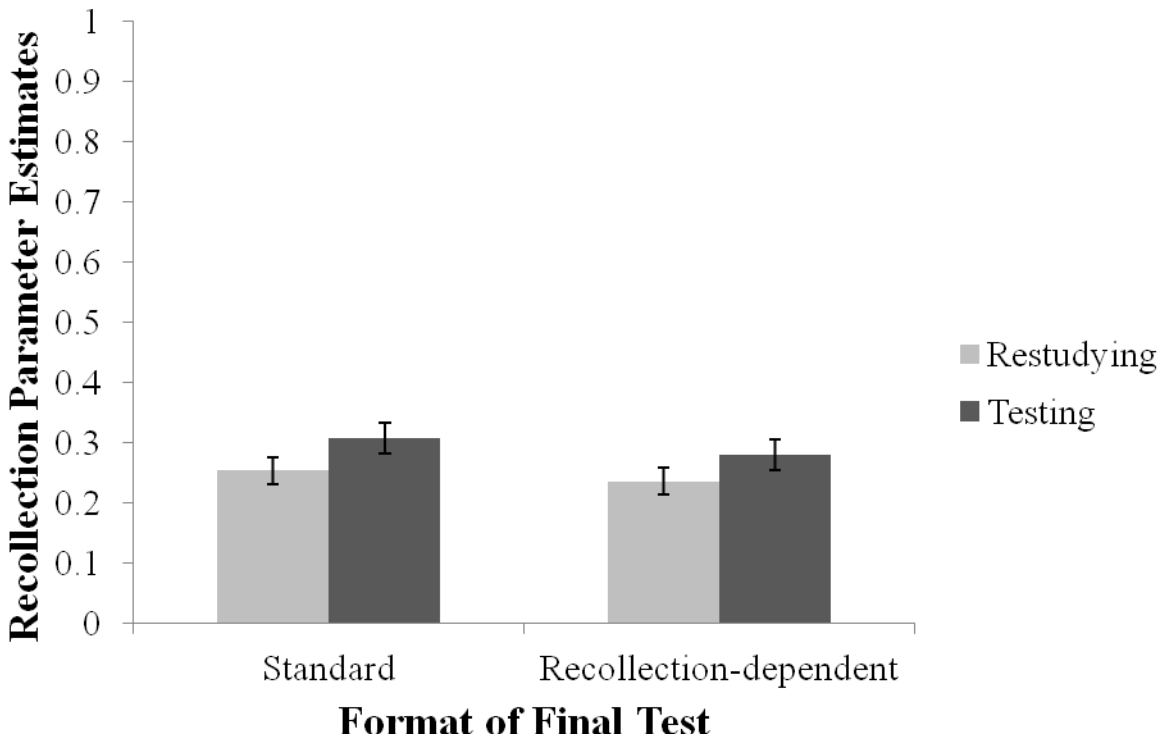

Figure 10. Recollection parameter estimates for old versus novel items for Experiment 2.

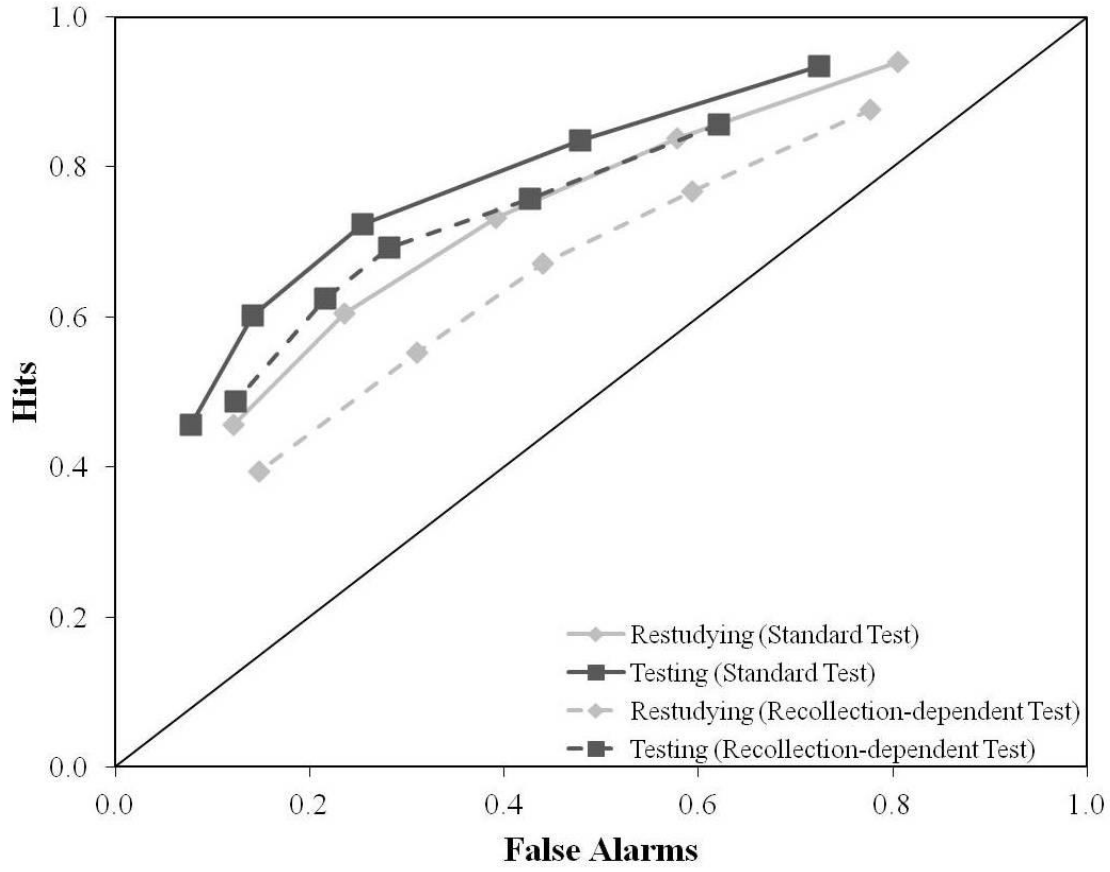

Figure 11. ROCs for old versus novel items for Experiment 2.

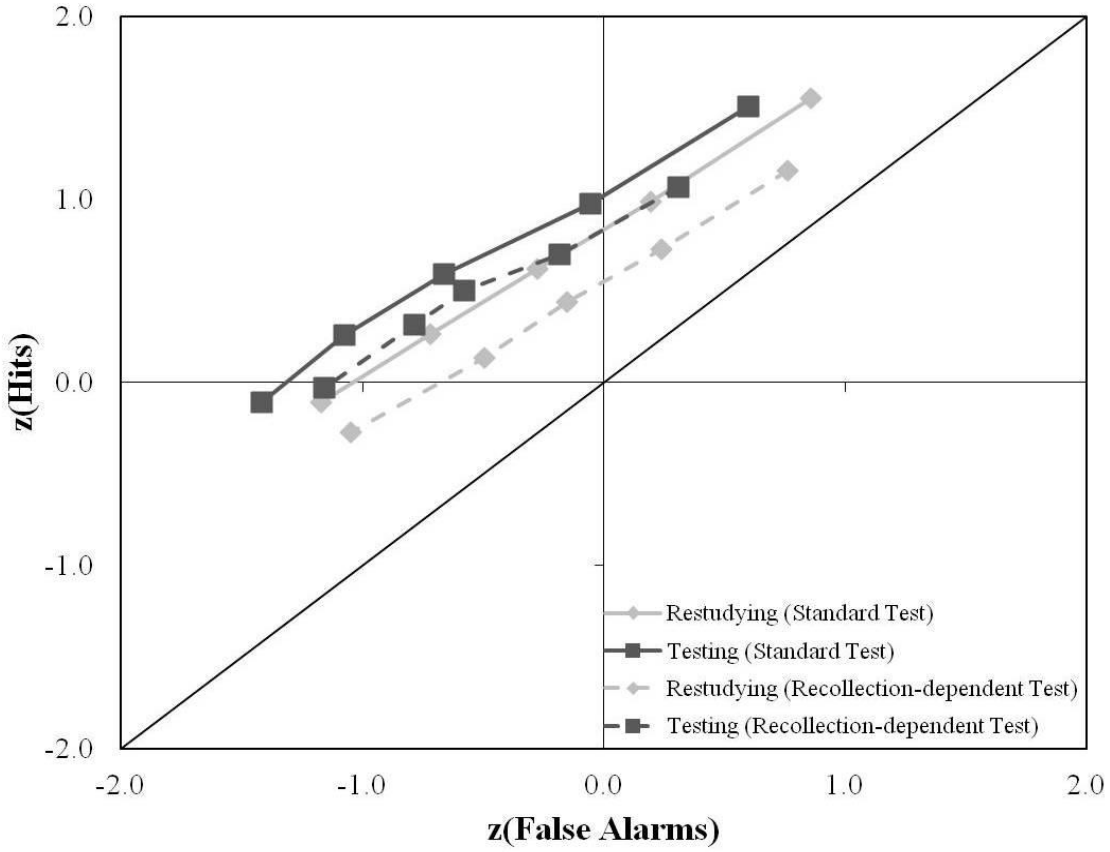

Figure 12. *z*ROCs for old versus novel items for Experiment 2.

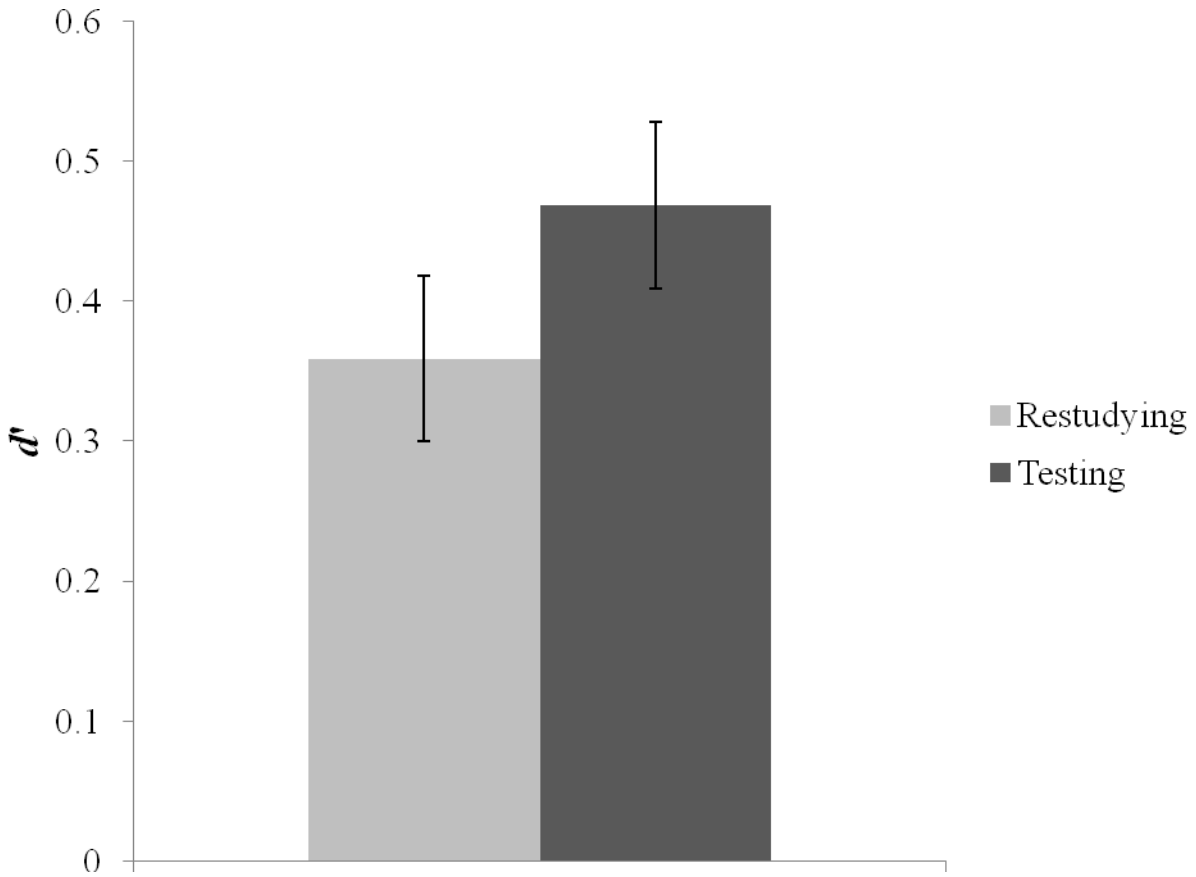

Figure 13. *d*' values for old versus plurality-reversed items for Experiment 2.

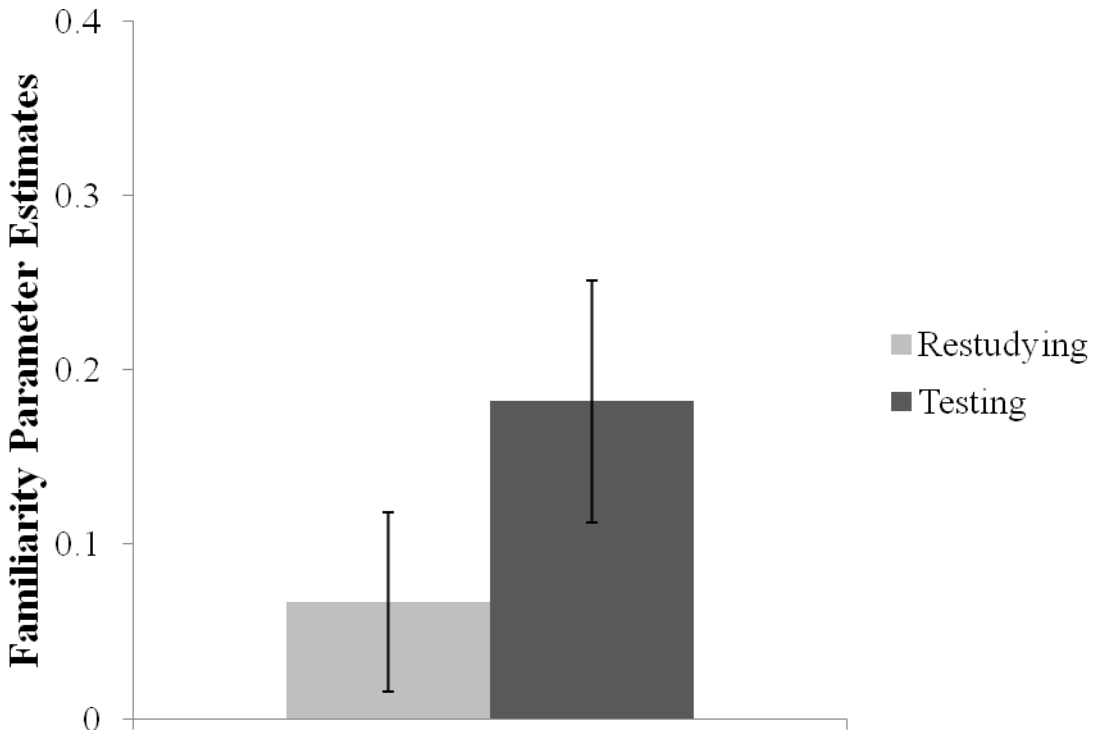

Figure 14. Familiarity parameter estimates for old versus plurality-reversed items for Experiment 2.

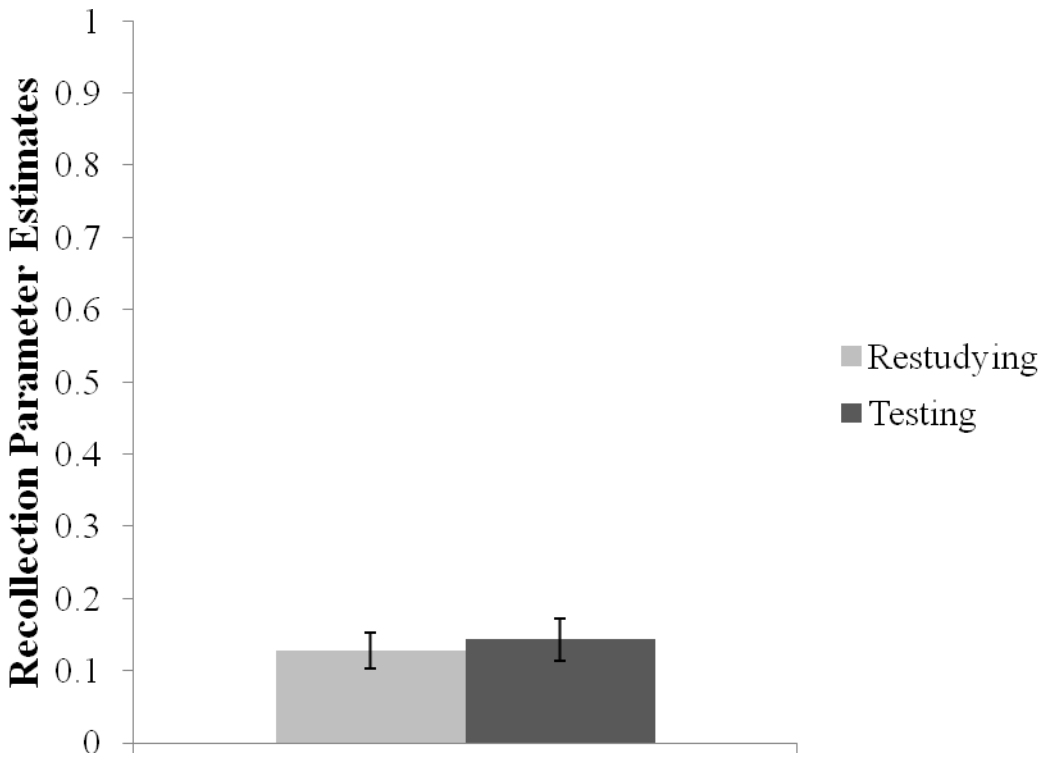

Figure 15. Recollection parameter estimates for old versus plurality-reversed items for Experiment 2.

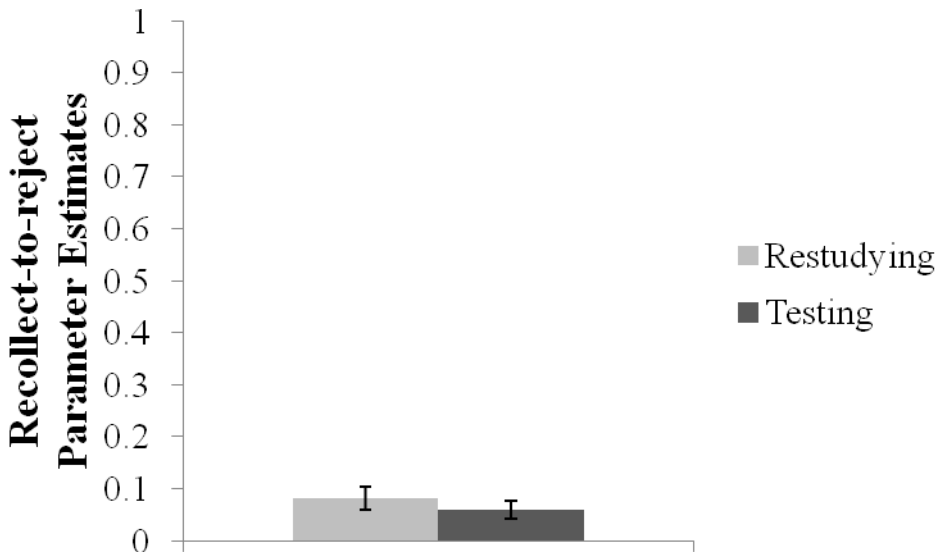

Figure 16. Recollect-to-reject parameter estimates for old versus plurality-reversed items for Experiment 2.

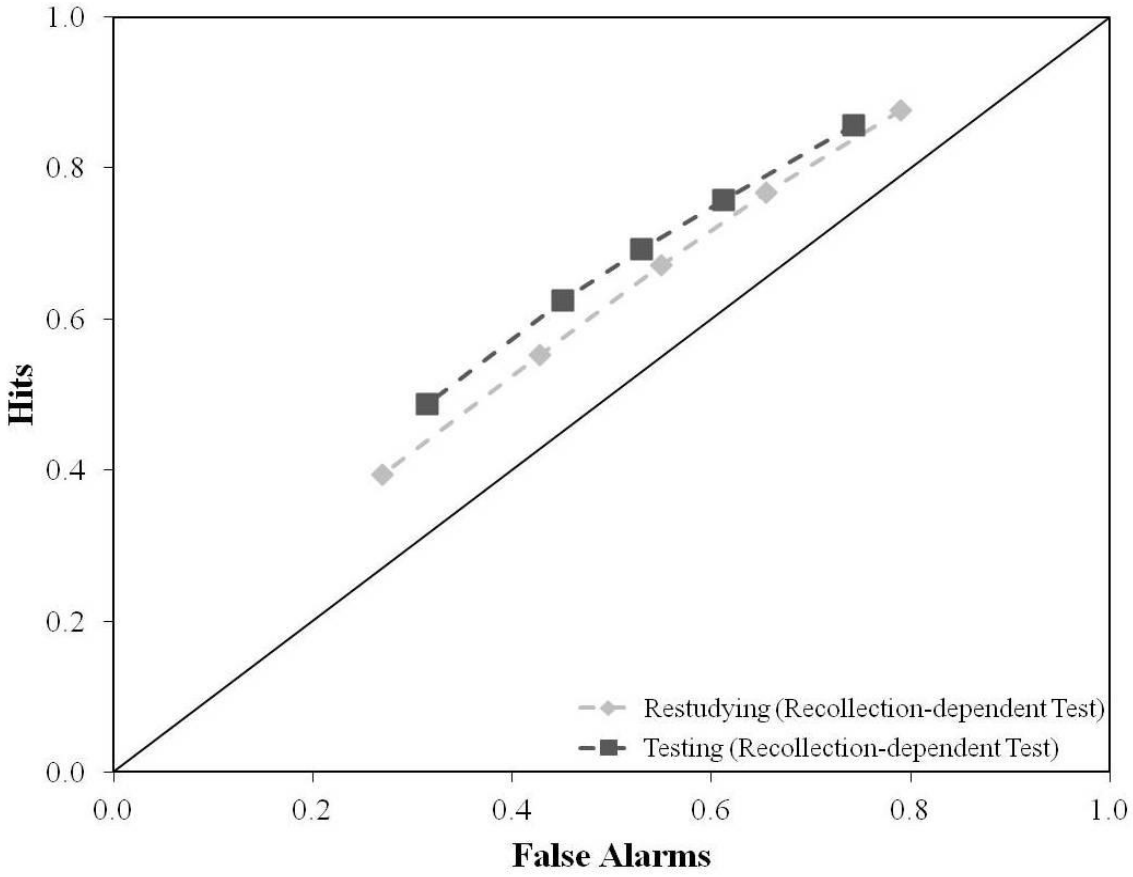

Figure 17. ROCs for old versus plurality-reversed items for Experiment 2.

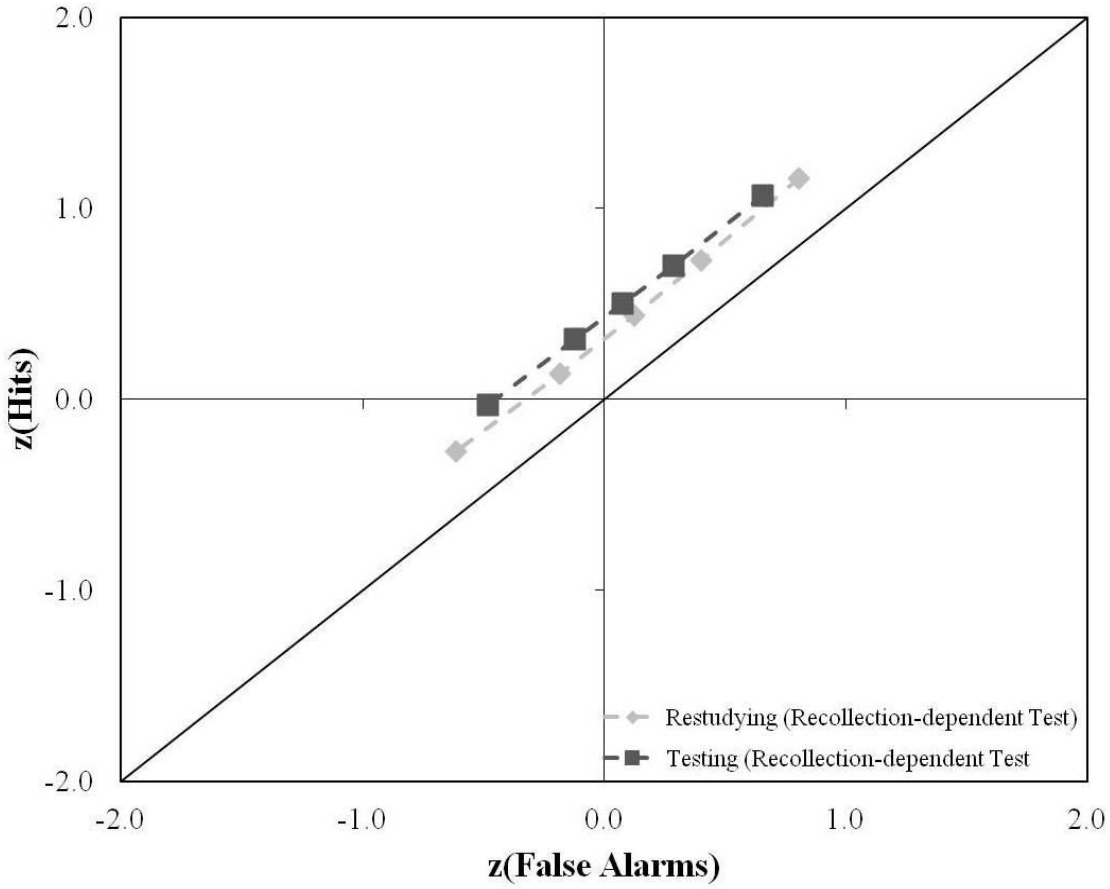

Figure 18. *z*ROCs for old versus plurality-reversed items for Experiment 2.

Appendix

Mean proportion of correct responses from the practice tests for Experiments 1 and 2.

	Practice Test 1	Practice Test 2	Practice Test 3
Condition	*M (SD)*	*M (SD)*	*M (SD)*
Exp. 1: One Practice Session	0.25 (0.10)		
Exp. 1: Three Practice Sessions	0.25 (0.13)	0.40 (0.17)	0.48 (0.19)
Exp. 2: Standard final test	0.29 (0.07)		
Exp. 2: Recollection-dependent final test	0.25 (0.09)		

VITA

Graduate College
University of Nevada, Las Vegas

Nicole J. Bies-Hernandez

Degrees:
Bachelor of Science, Psychology, 2006
Fayettevile State University

Master of Arts, Psychology, 2008
Fayetteville State University

Special Honors and Awards:
Fellowship, Summer Institute in Cognitive Neuroscience, University of California, Santa Barbara, June 24-July 7, 2012
2012 Part-Time Instructor Award for Social Sciences, School of Liberal Arts and Sciences, Nevada State College
College of Liberal Arts Dean's Graduate Student Stipend Award, University of Nevada, Las Vegas, Summer 2012
Graduate Student of the Year, Fayetteville State University, 2008

Publications:
Bies-Hernandez, N. J. (2012). The effects of framing grades on student learning and preferences. *Teaching of Psychology, 39(3)*, 176-180.

Schroeder, P. J., Copeland, D. E., & Bies-Hernandez, N. J. (2011). The influence of story context on a working memory span task. *Quarterly Journal of Experimental Psychology, 65(3)*, 488-500.

Copeland, D. E., Gunawan, K., & Bies-Hernandez, N. J. (2011). Source credibility and syllogistic reasoning. *Memory and Cognition, 39(1)*, 117-127.

Dissertation Title: Examining the Testing Effect using the Dual-Process Signal Detection Model

Dissertation Examination Committee:
Chairperson, David E. Copeland, Ph.D.
Committee Member, Mark H. Ashcraft, Ph.D.
Committee Member, Colleen M. Parks, Ph.D.
Committee Member, Joel S. Snyder, Ph.D.
Graduate College Representative, CarolAnne M. Kardash, Ph.D.